Inauguraldissertation
zur Erlangung des akademischen Doktorgrades (Dr. phil.)
im Fach Psychologie
an der Fakultät für Verhaltens- und Empirische Kulturwissenschaften
der Ruprecht-Karls-Universität Heidelberg

Titel der publikationsbasierten Dissertation

Nature and Validity of Complex Problem Solving

vorgelegt von
Dipl. Psych. Sascha Wüstenberg

Jahr der Einreichung
2013

Dekan: Prof. Dr. Klaus Fiedler
Berater: Prof. Dr. Joachim Funke und Dr. Samuel Greiff

Acknowledgements

First of all, I want to thank very warmly my two supervisors Prof. Dr. Joachim Funke and Dr. Samuel Greiff for their helpful support in the past three years. Joachim inspired me to think out of the box and to be open minded for input from other psychological and non-psychological sources. Samuel considerably improved my scientific writing and my ability to give talks. Both Samuel und Joachim encouraged me to present my research at conferences and to establish and maintain scientific contacts. Thereby, I was introduced to an activity called *scientific tourism*, combining work and leisure time at conferences all over the world. I still enjoy remembering our stays in Szeged, New Orleans, Cape Town, and all the other places we went to.

Further, I am deeply indebted to Dr. Jean-Paul Reeff who made it possible for me to write my dissertation on this project. I am grateful to Gyöngyver Molnár, Krisztina Tóth, and Benö Csapó from the University of Szeged, Hungary, for applying MicroDYN to Hungarian students. It was a great pleasure to work with such excellent people in an international cooperation. In this respect, it was also amazing to be part of the team that developed items for the measurement of problem solving in the Programme for International Student Assessment (PISA) in 2012.

I want to express my gratitude to my friends and colleagues Dr. Daniel Danner, Andreas Fischer, Jonas Müller, André Kretzschmar, Ursula Pöll, and Julia Hilse. It was great to have you around during my time as a doctoral student. I would also like to extend special thanks to all of our research assistants and students who took part in the studies.

Finally, I am very grateful to my whole family for their continuous support and I am deeply thankful to Chrissi for her love, which is invaluable to me.

Summary

This thesis investigates the nature and validity of complex problem solving (CPS). The main focus lies on analyses of three research questions dealing with CPS' (1) internal structure, its (2) structural stability combined with comparisons of performance differences across groups, and its (3) construct validity. In previous research, results on CPS' (1) internal factor structure have been solely based on samples with high cognitive performance, (2) structural stability of CPS across groups has not been tested yet, and analyses of performance differences across groups have been rather scarce. Further, results on (3) construct validity dealing with the relation of CPS to other measures of cognitive performance are inconsistent.

By applying a multiple task approach called MicroDYN to measure CPS, (1) internal factor structure is tested in groups of high school students, university students, and blue-collar workers, varying considerably in age and cognitive performance. Thereby, the interplay of theoretically derived CPS dimensions (*use of strategies*, *knowledge acquisition*, *knowledge application*) and their relation to characteristics of CPS tasks is analyzed in-depth. Further, for the first time in CPS research, (2) structural stability of CPS across groups is evaluated before performance differences are compared. Finally, (3) construct validity of CPS is investigated focusing on CPS' incremental validity beyond reasoning in explaining school performance. In summary, the present work addresses several gaps in existing research.

In chapter 1, the construct CPS and the measurement approach MicroDYN are introduced. Subsequently, previous results on internal structure, structural stability, performance differences, and construct validity are reported and followed by a brief description of the four empirical papers that are the main body of this thesis. The full papers are located in chapters 2 to 5. Whereas the first two papers are already in press after having successfully passed peer-review, the latter two papers are currently under revision.

Furthermore, chapter 1 and chapter 6 refer to other papers including supplementary contributions of the author of this thesis on CPS research, which are listed as “additional papers” on page 10.

The first empirical paper included in this thesis analysed internal structure and construct validity of CPS using data from a sample of university students. A 2-dimensional structure comprising the dimensions of *knowledge acquisition* and *knowledge application* fitted the data best and *use of strategies* as third dimension did not yield important information on CPS performance beyond *knowledge acquisition*. Further, CPS showed incremental validity beyond reasoning in explaining variance in school grades (cf. chapter 2).

The second paper investigated structural stability and performance differences in CPS across high school students of different ages in a Hungarian sample and related performance in CPS to reasoning and parental education. Measurement invariance analyses based on structural equation models confirmed structural stability of CPS across groups and revealed that CPS performance increased with higher age. Further, CPS showed incremental validity beyond reasoning in analyses based on manifest variables. Moreover, parental education predicted performance in CPS (cf. chapter 3).

The third contribution expands analyses of the second paper by comparing the structure of CPS across samples of high school students, university students, and blue-collar workers varying considerably in age and cognitive performance. As expected, CPS was measured invariant and participants in the academic track (university students, senior high school students) performed significantly better than participants in a non-academic track (blue-collar workers, junior high school students; cf. chapter 4).

The fourth paper investigated structural stability across gender and nationality by comparing German and Hungarian high school student samples. CPS was measured invariant and analyses on performance differences showed that Germans outperformed Hungarians and

males outperformed females. However, subsequent analyses revealed that performance differences could be partly explained by Hungarian females making less use of an efficient strategy (i.e., vary-one-thing-at-a-time) to generate knowledge. Further, influence of *use of strategies* as prerequisite for performance in CPS is discussed (cf. chapter 5).

Chapter 6 provides a general discussion of this research. Papers consistently yielded a 2-dimensional internal structure of CPS comprising the dimensions *knowledge acquisition* and *knowledge application*. CPS was measured invariant across samples varying in age, gender, and nationality, which is a prerequisite for comparing competency across groups. Analyses of performance differences not only showed that participants with higher education performed better in CPS, but also deepened the understanding of the influence of *use of strategies* as prerequisite of CPS performance. Further, encouraging results on incremental validity of CPS beyond reasoning were found. After this summary of results, strengths of the papers are outlined and shortcomings combined with an outlook for future research are provided. In this respect, five issues are tackled: the operationalization of CPS, its convergent and criterion validity, relation of CPS to general mental ability, use of process data, and trainability of CPS. In summary, this thesis advances knowledge about CPS and emphasizes its usefulness as an indicator of cognitive performance in addition to traditional measures.

Contents

Publication list for this cumulative dissertation

Paper 1:

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex Problem Solving − More than reasoning? *Intelligence, 40*, 1-14.

Paper 2:

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (in press). Complex Problem Solving in Educational Settings – something beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity. *Journal of Educational Psychology*.

Paper 3:

Greiff, S., & Wüstenberg, S. (submitted). Assessment with microworlds: factor structure, invariance, and latent mean comparison of the MicroDYN test. *European Journal of Psychological Assessment.*

Paper 4:

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (submitted). Determinants of cross-national gender differences in complex problem solving competency. *Learning and Individual Differences.*

Additional Papers

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitschke, A., & Funke, J. (2012). Die Bedeutung übergreifender kognitiver Determinanten für die Bewältigung beruflicher Anforderungen. Untersuchung am Beispiel dynamischen und technischen Problemlösens. *Zeitschrift für Erziehungswissenschaft, 15*(2), 363-391.

Danner, D., Hagemann, D.,Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., & Funke, J. (2011). Measuring Performance in Dynamic Decision Making. *Journal of Individual Differences*, *32,* 225-233.

Greiff, S., Holt, D. V., Wüstenberg, S., Goldhammer, F., & Funke, J. (submitted). Computer-based assessment of Complex Problem Solving: Concept, Implementation, and Application. *Educational Technology and Development.*

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement, 36*(3), 189-213.

Schweizer, F., Wüstenberg, S., & Greiff, S. (in press). Do Complex Problem Solving dimensions measure something beyond working memory capacity? *Learning and Individual Differences.*

1

Introduction

1.1 Introduction

Tell me and I forget,

Teach me and I remember,

Involve me and I learn.

Benjamin Franklin (1706 – 1790)

Acquiring and applying knowledge is a key competency for being successful in life. However, it is not only existing factual knowledge that helps to cope with daily problems that are encountered, but also the ability of a person to actively engage with a problem situation to generate new knowledge necessary to solve a problem (Funke, 2003; Raven, 2000). This competency of acquiring and applying knowledge while interacting with a problem situation is captured in the construct of *Complex Problem Solving* (CPS; Funke, 2001). Currently, CPS competency draws a lot of attention by educationalists and is regarded as an important key qualification that students should obtain during school education. In the framework of the Programme for International Student Assessment (PISA) developed by the Organisation for Economic Co-operation and Development (OECD), in which CPS competency was measured worldwide in the 2012 cycle, it is stated that "the acquisition of increased levels of problem solving competency provides a basis for future learning, for effective participation in society and for conducting personal activities." (OECD, 2010, p.7)[1]

Despite the awakening public interest, several important aspects about the nature and validity of CPS are yet to be addressed. As will be outlined, (1) the internal factor structure of CPS has only been tested in samples with high cognitive performance, that is, university students (e.g., Bühner, Kröner, & Ziegler, 2008), high school students attending tracks preparing for university (e.g., Kröner, Plass, & Leutner, 2005), or trainees for management positions (Wagener, 2001). However, internal structure has not been tested using several

[1] In the framework, the term *interactive problem solving* is used instead of complex problem solving. However, both labels refer to the same theoretical background (cf. OECD, 2010, p. 43).

heterogeneous samples varying considerably in cognitive performance. Further, (2) structural stability of CPS across samples varying in gender, age, or nationality has never been investigated and findings on performance differences across these groups are rather scarce. (3) Finally, results on incremental validity of CPS beyond other measures of general mental ability are inconsistent.

Thus, in this thesis, the following research questions are addressed by adapting an existent CPS test for application to various large samples consisting of junior and senior high school students, university students, and blue-collar workers:

Ad (1) Internal structure: Can the internal factor structure of CPS found in samples with high cognitive performance also be replicated in samples with less cognitive performance? And how are CPS dimensions interrelated?

Ad (2) Structural stability and performance differences: Can the construct CPS be measured equally well across different groups varying in gender, age, or nationality? How does performance in CPS differ across these groups? And, if there are differences, what are the reasons?

Ad (3) Construct Validity: How is CPS empirically related to other constructs, especially to other measures of general mental ability such as reasoning? Does CPS explain additional variance beyond other indicators of general mental ability in important criteria of life success?

All aspects are incorporated in the empirical papers (chapters 2 to 5). Prior to that, the construct CPS and measurement approaches to assess CPS competency are presented in section 1.2. Afterwards, the MicroDYN approach (Greiff, 2012) aimed at measuring CPS is introduced along with modifications conducted for applying MicroDYN in samples with fewer cognitive abilities in section 1.3. Preceding results and remaining questions on internal

structure, structural stability, performance differences, and construct validity are reported in section 1.4 followed by a brief description of the empirical papers in section 1.5.

1.2 CPS and its measurement

Imagine you have just bought your first ebook reader and you have no experience with such a device. If you do not want to read the manual, you will try to find out how it works by merely toggling through various menus, thereby generating knowledge about the menus' structure. After having explored the reader for a while, you will be able to intentionally adjust the settings (e.g., change fonts; add books to your virtual library, etc.). This is a typical situation considered as a complex problem involving dynamic interaction with a yet unknown system.

In such CPS tasks, no obvious method of solution is available and barriers between the initial state (e.g., ebook reader is turned off) and the goal state (e.g., reading a certain book) have to be reduced by applying non-routine cognitive activities (Funke, 2012; Mayer, 1992; Mayer & Wittrock, 2006). Although "the major problem of current research [on CPS] is the lack of a firm theory about dealing with complex problems" (Funke, 2012, p.685), it is widely acknowledged that problem solvers face two main demands: the need to generate knowledge about the systems' structure, *knowledge acquisition*, and the need to reach a certain goal by applying knowledge gathered beforehand, *knowledge application* (Funke, 2001). While acquiring and applying knowledge in a CPS task, problem solvers build a problem representation and derive a problem solution, which are the two major components of the problem solving process accountable to all kinds of problem solving (Mayer, 2003; Mayer & Wittrock, 2006; Novick & Bassok, 2005).

In contrast to simple static problems, CPS tasks usually contain many variables (aspect: complexity) that are highly interrelated (connectivity) and change dynamically (dynamics). There can be multiple, eventually contradicting goal states (polytelic situation) and the

problems' structure is opaque to participants (intransparency) (Funke, 2010). The task characteristics (1) complexity, (2) connectivity, (3) dynamics, (4) polytelic situations, and (5) intransparency can be allocated to competencies required by the problem solver. According to Funke (2001), (2) connectivity and (3) dynamics can be directly related to the competency of *knowledge acquisition* and *knowledge application.* That is, interrelated variables as indicators of connectivity require building an adequate mental model in *knowledge acquisition* and dynamics strongly influence system control attributed to *knowledge application*, but dynamics also have to be considered while acquiring knowledge about the systems' structure.

The other characteristics of a CPS task, (1) complexity, (4) polytelic situation, and (5) intransparency, can be allocated to competencies that may be considered as subdimensions of *knowledge acquisition* and *knowledge application*. (1) Complexity requires participants to reduce information by focussing only on important variables captured in the dimension *information reduction* and (4) polytelic situations call for the ability to weight and coordinate different goals measured in the dimension *evaluation* (Fischer, Greiff, & Funke, 2012; Funke, 2001). *Information reduction* may occur during *knowledge acquisition* and *knowledge application* as choosing relevant variables is important for both aspects, whereas *evaluation* can be assigned to *knowledge application*, because goal coordination and priority setting is more relevant when actually trying to reach goals than in generating knowledge. *Information reduction* and *evaluation* were mostly investigated from an experimental perspective (Blech & Funke, 2010; Haider & Frensch, 1996). An attempt to assess these dimensions psychometrically is currently under development and theoretically described by Fischer et al. (2012), but not part of this thesis.

Due to (5) intransparency, participants have to search for information to acquire knowledge. Funke (2001) stated that the ability of implementing appropriate strategies while searching for information may yield additional information on participants' cognitive

activities. This part of CPS competency is subsequently referred to as *use of strategies*, which is regarded as a predictor for *knowledge acquisition* in research on CPS, although sometimes labelled differently (e.g., information retrieval in Greiff, Wüstenberg, & Funke, 2012; rule identification in Kröner et al., 2005). It has to be noted that within this thesis, the dimension *use of strategies* solely captures the competency of implementing useful strategies to *acquire* knowledge, which can be distinguished from strategies that are used to *apply* knowledge already gathered (Schoppek, 2004). The latter kind of strategy is not considered in this thesis, but provides an interesting venue for future research (cf. chapter 6).

In summary, the focus of this thesis lies on the measurement of *use of strategies* (i.e., strategies to *acquire* knowledge), *knowledge acquisition* and *knowledge application*. For more information on theoretical aspects of problem solving refer to Funke (2003, 2012) as well as Mayer and Wittrock (2006).

1.2.1 Measurement of CPS

Within CPS research, two different measurement approaches can be distinguished that are based on either semantically rich scenarios, or formally constructed artificial tasks (Funke, 2010). Semantically rich scenarios were used to simulate the complexity of the real world as exactly as possible by implementing a huge amount of interconnected variables (Funke & Frensch, 2007). Examples are the CPS task *Tailorshop*, in which participants have to maximize the company value of a tailor manufactory (Funke, 2001), or *Moro*, which models the living conditions of a semi-nomad tribe (Strohschneider & Güss, 1999).

Semantically rich scenarios “represent a compromise between experimental control and realism” (Gonzalez, Vanyukov, & Martin, 2005, p. 274), posing two serious problems: (1) Due to the realistic context, genuine CPS competency is confounded with prior knowledge. The more prior knowledge a participant has (e.g., about managing a company), the less genuine CPS competency has to be applied to solve the respective task. For instance,

participants will not have to search for information to acquire new knowledge, if they already know the problem's structure. Consequently, participants' competency in *use of strategies* cannot be adequately measured. Thus, in semantically rich scenarios, performance is not clearly attributable to CPS competency, but to a mixture of CPS competency and prior knowledge (Süß, 1996). The influence of prior knowledge on problem solving performance is well-known and vividly illustrated in research on experts and novices (e.g., Sweller, 1988) and in experiments solely based on novices (Funke, 1992). In a recent study, Greiff, Holt, Wüstenberg, Goldhammer, and Funke (submitted) also showed that participants performed significantly worse in a task including relations between variables that did not match reality, than others who solved an otherwise identical task with variable relations corresponding to real world experiences. In summary, to measure genuine problem solving competency, the influence of prior knowledge on performance in a CPS task has to be reduced (cf. Paper 1).

(2) The second problem associated with semantically rich scenarios is related to the fact that in most scenarios one task with a specific system configuration is used to determine performance, leading to dependent indicators (cf. Paper 1). For instance, in *Tailorshop*, participants first explore the *Tailorshop* scenario for several months (i.e., rounds) to acquire knowledge, before they have to apply that knowledge in a control task lasting 12 months, in which the aim is to increase company value. In earlier research on *Tailorshop*, the total company value acquired at the end of the last control round was proposed as the indicator (cf. Funke, 1983). However, rounds within *Tailorshop* are not experimentally independent, because each action in a given round influences the status of the company value in the next round, which violates the assumption of uncorrelated errors (Danner, Hagemann, Holt, Hager, Schankin, Wüstenberg, & Funke, 2011). Recently, Sager, Barth, Diedam, Engelhart, and Funke (2011) provided a mathematical optimization model that expresses the best possible solution for each subsequent round given the decisions a participant already made before. Thus, participants' actions in each round can be compared to the optimal solution that is still

possible considering the previous interventions (Sager et al., 2011). Although applying this procedure leads to a more objective evaluation of participants' performance gained by the comparison with mathematically optimal solutions, the core problem is not solved. The performance in each *Tailorshop* round is dependent on the knowledge that the participant acquired beforehand while exploring the scenario (as well as from prior knowledge gathered outside the test situation). In other words, although it is possible to quantify the quality of *knowledge application* within a given round, these measures are dependent on the same *knowledge acquisition* phase. This type of problem is referred to as one-item-testing (e.g., Paper 1).

In summary, the problems described can be solved by minimizing influence of prior knowledge and by measuring CPS competency with multiple, independent tasks that all include *knowledge acquisition* and *knowledge application* phases. Suitable formalisms to integrate both aspects are artificially constructed CPS tasks based on either finite state automata (FSA) including qualitative relations between variables (e.g., modelling a ticket vending machine), or linear structural equations (LSE) incorporating quantitative relations (e.g., different amount of training related to the properties of a handball team). For a detailed description of LSE and FSA see Funke (2001). However, in most scenarios based on FSA (e.g., *Space Shuttle*, Wirth & Funke, 2005) and LSE (e.g., *MultiFlux*, Kröner et al., 2005) only single tasks were applied. Multiple tasks with varying causal structure were only used in between-subjects designs and not in within-subjects designs (e.g., *ColorSIM*, Kluge, 2008).

Only recently has an ambitious approach labelled MicroDYN (Greiff, 2012) emerged. This approach (1) minimizes the influence of prior knowledge and (2) uses multiple, independent LSE tasks. The semantic embedding used in MicroDYN includes variables that are labelled either fictitiously or without any deep semantic meaning (e.g., Training A, B, C for different sorts of training). This implies that tasks can be solved without domain-specific

prior knowledge about the semantic embedding allowing an unconfounded assessment of CPS competency (Funke, 2001; cf. Paper 1). Further, multiple independent tasks with only few time-on-task (i.e., about 5 minutes in total) are applied to enable a psychometrically adequate assessment of CPS with independent indicators of performance (Greiff, 2012; Sonnleitner et al., 2012).

This thesis was part of a project, in which the main focus was to develop and expand MicroDYN in a way that performance in CPS can be measured in a reliable and valid manner across different target populations. These target populations vary in cognitive abilities, allowing researchers to test the main research questions on internal structure, structural stability, performance differences, and construct validity of CPS. Therefore, in the next section, the first version of MicroDYN as well as preliminary empirical results are described and followed by a summary of changes that were conducted to achieve the project goals.

1.3 MicroDYN

MicroDYN is based on minimal complex LSE-tasks (see Figure 1), extending the DYNAMIS approach (Funke, 1993) by incorporating multiple tasks.

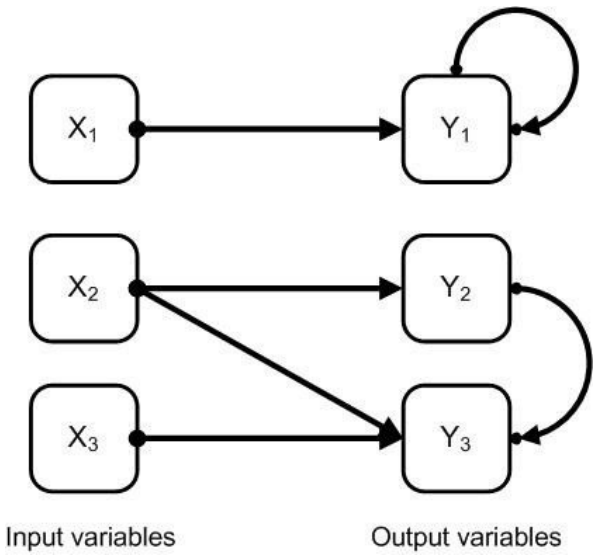

Figure 1. Example of the structure of a typical MicroDYN system displaying input variables (X_1, X_2, X_3) and output variables (Y_1, Y_2, Y_3).

Although MicroDYN has passed through several iterations during its development, the general procedure remains the same. While working on MicroDYN, problem solvers are confronted with two different phases: First, they have to apply appropriate strategies (dimension: *use of strategies*) to find out how input variables are related to output variables. Within this process, they generate knowledge about the causal structure of the task, which they have to represent in sort of causal diagrams in a situational model (dimension: *knowledge acquisition*). Second, participants have to reach certain target values for the output variables by applying the knowledge acquired beforehand (dimension: *knowledge application*).

The first version of MicroDYN was based on a set of tasks using a scenario of a chemistry laboratory (see Figure 2).

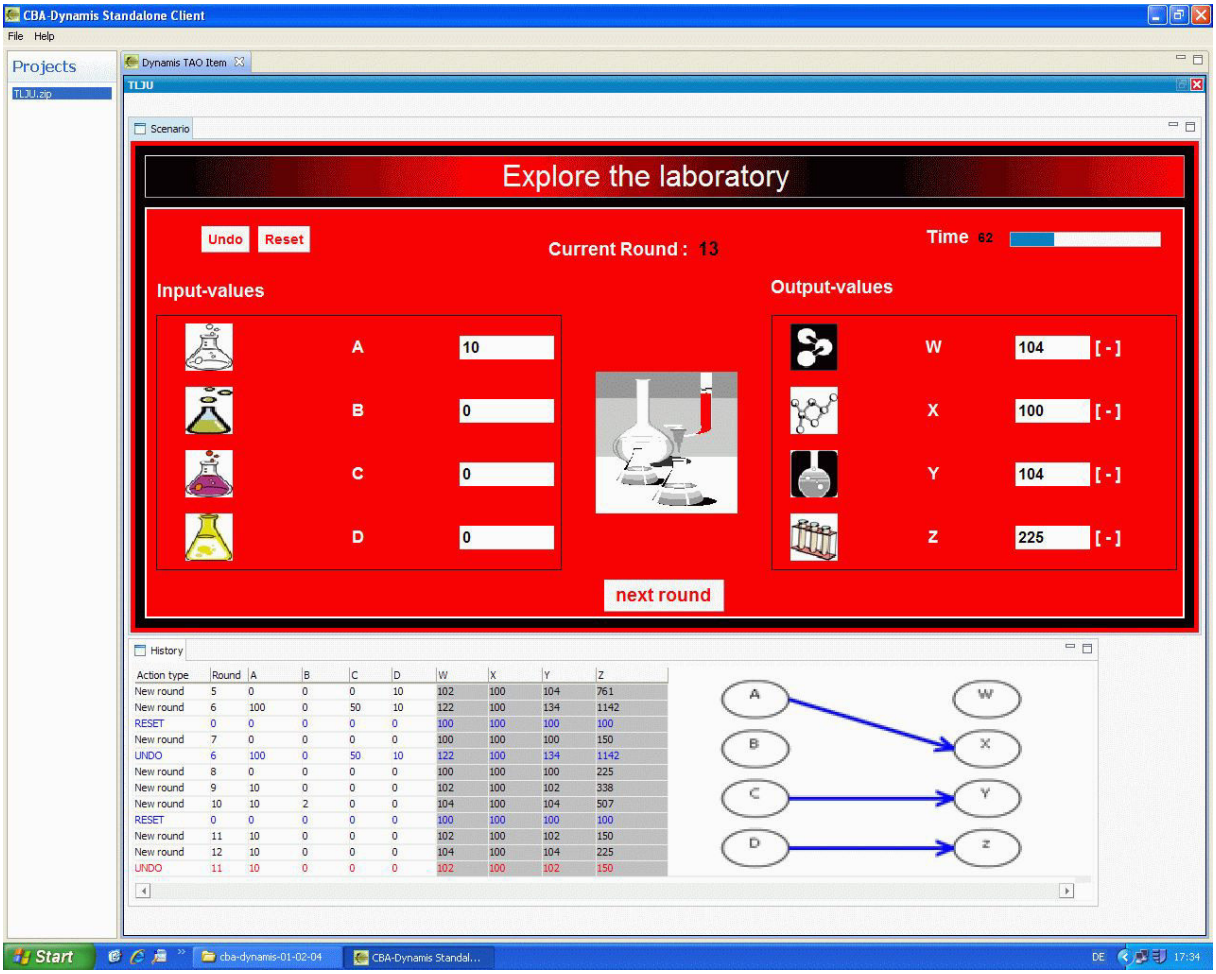

Figure 2. Screenshot of an early MicroDYN-task with four input variables A, B, C, and D, and four output variables W, X, Y, and Z.

Within the first phase of this task, participants have to find out how four input variables (labelled A, B, C, D) are related to four output variables (labelled W, X, Y, Z) by using appropriate strategies. Thereby, participants vary input variables by inserting numerical values (e.g., value "10" for variable A; see Figure 2). After clicking Next Round, system states change according to participants' actions and dynamics inherent in the system. By interpreting changes in the numerical values of the four output variables, participants acquire knowledge about variable interrelations. This knowledge has to be transferred into a situational model by drawing paths between variables and by inserting corresponding path weights. In the second phase, the correct model is presented to participants and they have to reach given target values at the output variables by correctly manipulating input variables. For *use of strategies*, participants' use of efficient strategies such as vary-one-thing-at-a-time strategy (VOTAT; cf. Greiff, 2012; Rollett, 2008) is evaluated to identify causal relations. For *knowledge acquisition,* the correctness of the situational model is scored and for *knowledge application*, the extent to which target goals are reached. A detailed description of the first version of MicroDYN including scoring procedures can be found in Greiff (2012) and Greiff et al. (2012).

In CPS research, it is theoretically assumed that the CPS dimensions *use of strategies*, *knowledge acquisition*, and *knowledge application* are empirically distinguishable, and that *use of strategies* is a prerequisite for successful *knowledge acquisition*, which in turn predicts performance in *knowledge application* (e.g., Kröner et al., 2005). Empirical results based on the first version of MicroDYN revealed that the dimensions are significantly related and that a 3-dimensional structure fitted well and even better than a 2-dimensional model or a 1-dimensional model (N=114; Greiff, 2012). This result could be replicated on another sample in a cross-validation study, in which the author of this thesis was involved (N=140; Greiff et al., 2012). Additionally, performance in *knowledge acquisition* and *knowledge application* in MicroDYN significantly predicted performance in *knowledge acquisition* and *knowledge*

application in the CPS task *Space Shuttle* (R^2=.43-.45, p<.01; Greiff et al., 2012), which requires participants to control a space shuttle and a space vessel (Wirth & Funke, 2005). Thus, MicroDYN showed convergent validity to *Space Shuttle* that was also used in the German national extension of PISA 2000 (Wirth & Klieme, 2003). Finally, *knowledge acquisition* in MicroDYN was significantly related to school grades (R^2=.24, p<.01), providing first results on concurrent validity.

In summary, although being preliminary, reported results imply that MicroDYN provides a fruitful approach to assess CPS competency. However, the specific test used in Greiff (2012) and Greiff et al. (2012) was designed for university students. In order to also capture CPS competency of test takers with lower cognitive performance, several adjustments on MicroDYN had to be carried out that are reported in section 1.3.1.

1.3.1 Adaptations of MicroDYN

In the previous version of MicroDYN, items were a bit too difficult even for university students, indicated by low relative frequencies of correct solutions in *knowledge acquisition* (p=.07-.67) and *knowledge application* (p=.04-.69) (Greiff et al., 2012). Several changes were introduced to make MicroDYN suitable for a broad range of participants varying in cognitive performance (see Table 1). Therefore, task difficulty and the user interface were adapted. As recommended by Greiff (2012), task difficulty was decreased by adjusting (1) task configuration. Thus, the number of effects between variables was lowered and only few tasks with indirect effects were used. According to Greiff (2012), an effect is labelled as indirect effect if either one output variable influences another output variable (also labelled side effect), or an output variable influences itself (also labelled eigendynamic). Further, identical path weights were applied across tasks and the number of input and output variables were decreased (cf. Table 1). The latter change also reduces time pressure, because fewer variables have to be considered while working on tasks.

Table 1

Overview of Differences between the Previous and the Current Version of MicroDYN

	Aspects changed	Previous Version of MicroDYN	Current version of MicroDYN with reduced task difficulty (applied in this thesis)
(1)	Task configuration	More tasks with indirect effects than tasks with only direct effects	More tasks with only direct effects than tasks with indirect effects
		High number of tasks with at least four effects between variables	Low number of tasks with at least four effects between variables
		Only tasks with 4 input variables and 4 output variables	Tasks with up to 3 input variables and up to 3 output variables
		Varying path weights across tasks	Identical path weights across tasks: Input variable on output variable (β=2.0) Output variable to another output variable (β=0.2) Output variable on itself (β=1.33)
(2)	Range of possible input values	Numerical values ranging from -100 to +100	Slider with 5 options ranging from -- (hidden value -2) to ++ (hidden value +2)
(3)	Causal model	Participants insert paths and path weights of relations	Participants insert only paths between variables without path weights
(4)	Display of output variables	Numerically	Numerically and graphically
(5)	Coverstories	One coverstory per test	One coverstory per task

Moreover, (2) the range of possible input values was restricted. Instead of inserting numerical values for the input variables, participants have to move sliders ranging from "--" to "++" (see Table 1 and Figure 3), which decreases the variety of possible interactions with the task environment and the possibility that participants use inappropriate input values leading to developments that cannot be adjusted for.

In the current version, participants only have to draw arrows between variables in the (3) causal model if they think they are related, but they do not have to implement concrete path weights. For instance, if participants think "Norilan" and "Fresh" are related, they only have to draw an arrow between both variables (see bottom of Figure 3), but do not have to enter the

concrete path weight (e.g., value "+2") as realized in the previous version (cf. Greiff, 2012). This modification decreases the influence of mathematical abilities on performance in *knowledge acquisition*. For the same reason, (4) input and output variables were displayed both numerically and graphically instead of only numerically. Finally, in order to enhance motivation of participants, (5) different coverstories were applied (e.g., perfume displayed in Figure 3; Handball training displayed in Paper 1).

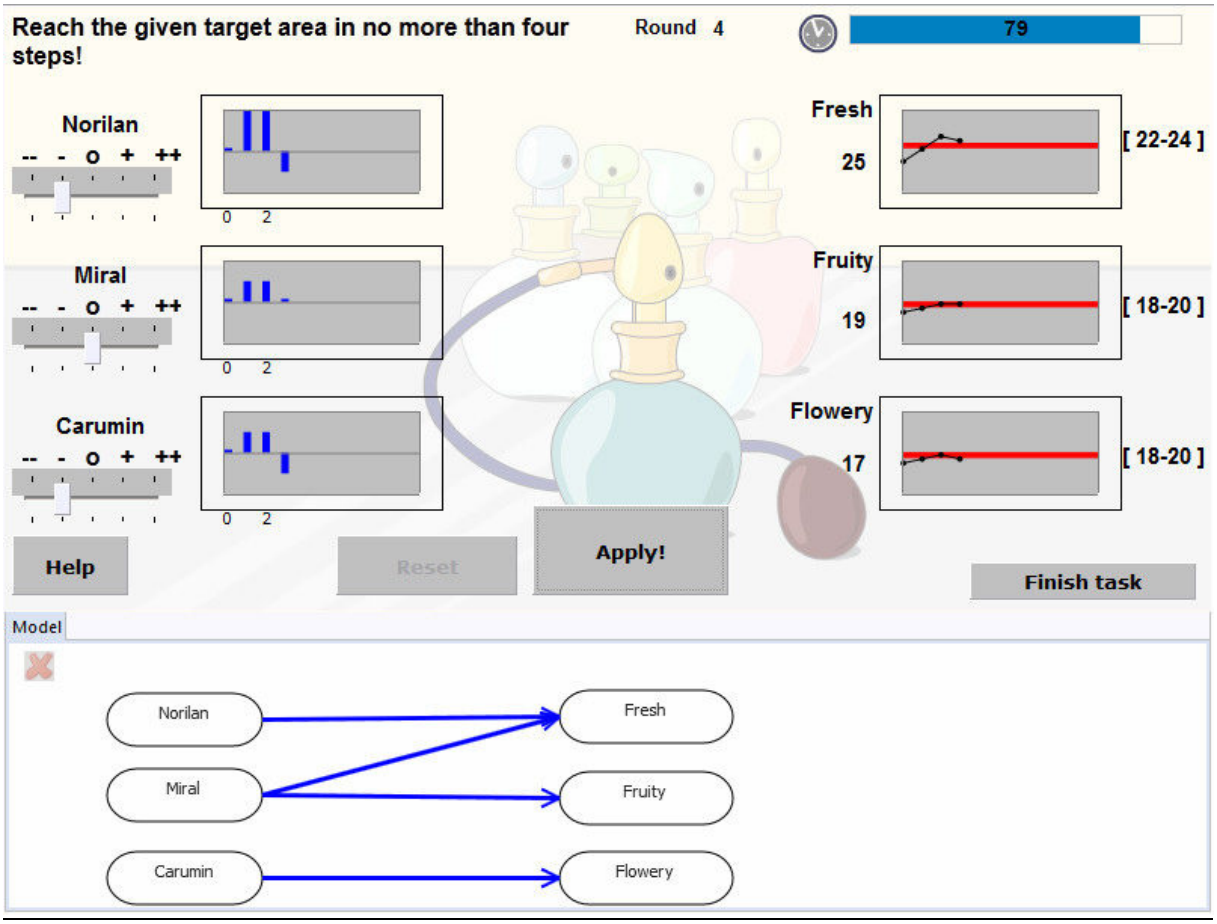

Figure 3. Screenshot of a new MicroDYN-task (Perfume) with three input variables Norilan, Miral, and Carumin, and three output variables Fresh, Fruity, and Flowery.

1.4 Previous empirical results on internal structure, structural stability, and construct validity of CPS

In the subsequent sections, previous empirical results on CPS' internal structure, structural stability, and construct validity based on other approaches are summarized and complemented with findings based on the first version of MicroDYN.

1.4.1 Internal structure of CPS

In nearly all empirical studies on CPS, the dimensions *knowledge acquisition* and *knowledge application* were significantly related (Bühner et al., 2008; Greiff et al., 2012; Kluge, 2008; Süß, 1996; Wittmann & Hattrup, 2004). Opposite results were only found if participants were forced to control a CPS task immediately without having the time to acquire knowledge about relations between variables beforehand (Berry & Broadbent, 1995; for further details see Kröner et al., 2005; Vollmeyer & Burns, 1996).

However, the most fundamental shortcoming of all studies reporting data on internal factor structure is non-representativeness of participants who were either university students (e.g., Bühner et al., 2008; Greiff et al., 2012; Kluge, 2008), senior high school students (e.g., Kröner et al., 2005; Süß, 1996), or trainees aspiring to management positions (Wagener, 2001), all having high cognitive abilities. Studies based on employees showing a broader range of cognitive performance did not mention results on the internal structure of CPS (e.g., Danner, Hagemann, Schankin, Hager, & Funke, 2011; Danner, Hagemann, Holt, Hager, Schankin, Wüstenberg, & Funke, 2011). In this respect, it has yet to be tested if the internal structure comprising the dimensions *knowledge acquisition* and *knowledge application* also holds in samples with lower cognitive abilities, such as high school students attending different school tracks or blue-collar workers. Replicating the internal structure in

heterogeneous samples is a requirement for evaluating structural stability and analysing mean differences, as will be outlined in section 1.4.2.

In addition to analyses on *knowledge acquisition* and *knowledge application*, some authors recommend to also analyze process data that can be used to score the dimension *use of strategies* obtaining more information on how participants interact with a CPS task (e.g., Funke, 2001; Kröner et al., 2005). *Use of strategies* is either investigated from an experimental perspective (Vollmeyer & Burns, 1996; Vollmeyer, Burns, & Holyoak, 1996), or by modeling it psychometrically as a separate dimension (Greiff, 2012; Greiff et al., 2012; Kröner et al., 2005). Kröner et al. (2005) were the first showing that three dimensions could be empirically distinguished on a latent level. However, their data relied on only one single CPS task, implying a dependency of indicators, which lead to a violation of test theoretical assumptions (cf. Paper 1). Despite this psychometrical shortcoming, the assumptions of three dimensions as proposed by Kröner et al. (2005) seem to be valuable to gain further information on the relation between *use of strategies*, *knowledge acquisition*, and *knowledge application*. As already outlined, first empirical results using the multiple-task approach MicroDYN in its first version also revealed that a 3-dimensional model with the dimensions *use of strategies*, *knowledge acquisition* and *knowledge application* fitted well (Greiff et al., 2012). However, due to several changes applied to MicroDYN (cf. section 1.3.1), results on internal structure cannot be adopted and analyses have to be done repeatedly. For instance, more mathematical ability was needed to solve MicroDYN tasks in the first version, in which path weights between input and output variables differed within and across tasks. This may have had different influences on *knowledge acquisition* and *knowledge application*. In *knowledge acquisition*, participants can interpret path weights subsequently to derive knowledge about the causal model, whereas in *knowledge application*, paths with differing path weights on the same output variable have to be considered simultaneously to reach target goals at output variables. Consequently, differing influence of mathematical ability (which is

needed more in *knowledge application*) may underestimate correlations between *knowledge acquisition* and *knowledge application* in the first version, influencing dimensionality of the construct. Nevertheless, both dimensions should be empirically distinguishable as they measure different aspects of CPS competency.

In going beyond previous research, this thesis not only relates the influence of *use of strategies* on *knowledge acquisition* and *knowledge application* (cf. Paper 1) and tests dimensionality of CPS, but also analyses reasons and prerequisites of results obtained and discusses generalizability (cf. Paper 4; chapter 6).

1.4.2 Structural stability and performance differences in CPS

As already mentioned, studies on CPS were conducted on high school students, university students, or employees, but studies combining all of these groups are non-existent. Thus, unlike research on other constructs (e.g., Wechsler Intelligence Scale for Children; Chen, Keith, Weiss, Zhu, & Li, 2010), the question if a measurement device assesses CPS competency equally well across different groups with varying demographical background has never been tested before. Measurement equivalence holds if the structure of the construct does not change across groups, which is usually evaluated by analysing measurement invariance (cf. Byrne & Stewart, 2006). Within this thesis, measurement invariance is tested across participants of different age groups (Paper 2; Paper 3), nationalities (Paper 4), and gender (Paper 4).

Evaluating measurement invariance has two advantages: First, it helps to answer important theoretical questions about the applied measurement device (i.e., the CPS test used). For instance, in MicroDYN, some coverstories that are embedded in a more male-oriented context (e.g., mixing chemical elements) were used and others include a more female context (e.g., mixing a perfume; cf. Figure 3). If measurement invariance holds, item difficulties are identical for both males and females, implying that a specific coverstory does not reward a

certain gender group. The same applies to differences across countries. By testing measurement invariance, cross-national differences in the structure of the construct can be revealed as well as errors occurring when tasks are translated from one language to another (Chen, 2008).

Second, only if measurement invariance holds, implying that the construct CPS does not change across groups, performance differences across these groups can be interpreted meaningfully (Byrne & Stewart, 2006). In the last few decades, analyses of performance differences in cognitive abilities across students (e.g., OECD, 2009) and adults (e.g., OECD, 2012) have increased substantially (McGehee & Griffith, 2001). In CPS research, however, studies on cross-cultural differences (Güss, Tuason, & Gerhard, 2010; Strohschneider & Güss, 1999) or gender differences (Wittmann & Hattrup, 2004) have been rather scarce and mostly more experimentally than psychometrically oriented (cf. Paper 4). To the best of the authors' knowledge, comparisons across age are non-existent (cf. Paper 2; Paper 3).

1.4.3 Construct validity of CPS

Research on CPS started because scientists such as Dörner (1986) criticized that traditional measures of general mental ability do not assess the competency to solve complex problems in real life, although they are often used as predictors of performance in different domains (e.g., Schmidt & Hunter, 2004). Although CPS and measures of general mental ability such as reasoning have some aspects in common, there are substantial differences (Raven, 2000; cf. Paper 1). For instance, CPS tasks require the ability of actively generating knowledge while interacting with a dynamic system (Raven, 2000; Funke, 2001), which is not measured in static tasks commonly used to assess general mental ability.

The predominant question is therefore how CPS is empirically related to well-established indicators of general mental ability. Due to parsimony reasons, separating the construct CPS from other measures of general mental ability is only meaningful if CPS

includes unique aspects contributing to the explanation of real world performance. This can be proven by investigating incremental validity of CPS in important performance criteria such as problem solving performance in daily life, or performance at work, or in school. If incremental validity of CPS could not be shown, the advantage of CPS as construct would be limited. Thus, this thesis investigates how CPS is related to a measure of general mental ability, analyses incremental validity beyond it and discusses how CPS can be integrated in the current state-of-the-art theory of general mental ability, Carroll's (1993, 2003) three stratum theory of intelligence (cf. Paper 1). In this theory, a latent factor of general mental ability is located at the third stratum, which captures common aspects of a few broad factors assumed to be at the second stratum (e.g., fluid and crystallized intelligence). Factors on Stratum 2 can be further divided into narrower aspects on the first stratum. As outlined in Paper 1, reliable and valid CPS tasks did not exist at the time Carroll conducted his analyses of empirical studies dealing with cognitive performance that were the foundation of his theory.

Empirical results on the additional value of CPS beyond existing measures of general mental ability were rather inconsistent. Whereas Süß (1996) as well as Wittmann and Hattrup (2004) stated that performance in CPS measured by the *Tailorshop* could be sufficiently explained by prior knowledge and measures of general mental ability (Berlin Intelligence Structure Test ─ BIS-K; Jäger, Süß, & Beauducel, 1997), Danner, Hagemann, Schankin, Hager, and Funke (2011) showed that CPS assessed by a latent CPS factor based on *Tailorshop* and the Advanced Progressive Matrices (APM; Raven, 1958) explained variance in supervisory ratings even beyond measures of general mental ability. As outlined in section 1.3, prior knowledge that is helpful in semantically rich scenarios may deteriorate genuine CPS performance, explaining some of the inconsistencies mentioned.

Finally, this thesis also investigates influence of parental education on both CPS and traditional measures of general mental ability in an exploratory analysis. Parental education is

considered the most important socioeconomic factor related to school performance (Myrberg & Rosen, 2008). Whereas small influences of parental education to g have already been found (Rindermann, Flores-Mendoza, & Mansur-Alves, 2010), relations of parental education and performance in CPS has never been investigated before. Good education helps parents to provide appropriate learning environments and cognitive stimulation for their children (Davis-Kean, 2005). Thus, children of well-educated parents may often be confronted with interactive situations, which are fundamental to acquiring and applying new knowledge, leading to significant relations between parental education and CPS. The younger the children are, the stronger the influence of parental education on childrens' cognitive development (Alexander, Entwisle, & Bedinger, 1994; Davis-Kean, 2005). Thus, as it includes samples of junior high school students aged 11 to 19 years, this thesis is suitable for exploring the relation of parental education and CPS (cf. Paper 2).

1.5 Main data collection and preview of papers

The papers included in this thesis are based on five large data collections (cf. Table 2) consisting of university student samples (collected in summer and autumn 2010), Hungarian students (collected at Hungarian schools by our colleagues at the University of Szeged in spring 2011), German junior high school students (spring 2011), German senior high school students (summer 2011), and blue-collar workers (collected at a big car manufacturer in winter 2011). Data collections are partly combined in the papers of this thesis as shown in Table 2, because one main research aim is to evaluate structural stability of CPS measured by MicroDYN across groups varying in age, gender, and nationality, and to compare their performance.

MicroDYN tasks used in the different samples are almost identical (cf. linear structural equations in Appendices of Paper 1 to Paper 4). When comparing performance across

different samples, only tasks that are completely identical for all groups are included in the analyses.

Table 2

Overview of the Five Data Collections Containing University Students, Hungarian High School Students, German Junior and Senior High School Students, and Blue-Collar Workers, and their Appearance in the Papers

	University students	**Hungarian high school students**	**German junior high school students**	**German senior high school students**	**Blue-Collar Workers**
	N=222	N=855	N=309	N=484	N=181
	-	grades 5 to 11	grades 8 to 10	grades 11 to 13	-
Paper 1 (Chapter 2)	N=222				
Paper 2 (Chapter 3)		N=855			
Paper 3 (Chapter 4)	N=222		N=309	N=484	N=181
Paper 4 (Chapter 5)		Only grades 8 to 11 N=479	N=309	Only grade 11 N=102*	
			Combined to one sample with grades 8 to 11 N=411		

* In Paper 4, only Grade 11 was chosen from the German senior high school sample to match samples collected in Germany and Hungary regarding grades. The German senior high school sample (*N*=102) used for these analyses was drawn from the same school as the German junior high school student sample.

The first paper investigates internal structure and construct validity. The main focus lies on incremental validity of CPS measured by MicroDYN beyond reasoning in predicting school grades. The second paper analyses structural stability and performance differences across high school students of different age in a Hungarian sample, and broadens analyses on

construct validity including the relation between CPS, reasoning, and parental education. The third paper expands analyses of the second paper by comparing the structure of CPS across samples varying considerably in age and cognitive performance. Finally, the fourth paper compares CPS performance of Hungarian and German students and investigates potential gender effects. Furthermore, the role of applying appropriate strategies while exploring a MicroDYN task is discussed in this last work.

1.5.1 Preview of Paper 1

Can three dimensions of CPS be empirically distinguished when using a new version of MicroDYN? Is CPS separable from reasoning, and, if so, does CPS explain additional variance in school grades beyond reasoning? These are the main questions this paper aims to address. Previous studies by Kröner et al. (2005) who used a CPS task based on linear structural equations similar to MicroDYN, showed that CPS can be modelled by a 3-dimensional structure and that CPS and reasoning are related. Concerning incremental validity, Danner, Hagemann, Schankin, Hager, and Funke (2011) reported that CPS measured by a semantically rich scenario *Tailorshop* (Funke, 2001), incrementally predicted supervisor ratings in a sample of employees even beyond measures of reasoning.

However, the present paper goes beyond the work of Kröner et al. (2005) and Danner, Hagemann, Schankin, Hager, and Funke (2011), because analyses on CPS are based on a multiple-task approach, having several psychometrical advantages such as independent indicators of performance (cf. Paper 1). Furthermore, research on CPS based on LSE tasks, in which genuine problem solving competency is not confounded with prior knowledge, has never investigated incremental validity of CPS beyond reasoning.

To enable readers to closely follow differences between this paper and the work of Kröner et al. (2005), dimensions of CPS in Paper 1 are named similar to Kröner et al. (2005) who chose to label *use of strategies* as *rule identification*, *knowledge acquisition* as *rule*

knowledge and *knowledge application* as *rule application*.[2] Whereas the terms used by Kröner et al. (2005) can also be found in research on inductive reasoning (cf. Babcock, 2002), *knowledge acquisition* while building a problem representation and *knowledge application* to derive a problem solution are unique to problem solving research (Funke, 2001; Novick & Bassok, 2005). The labels used within problem solving research and those applied by Kröner et al. (2005) include slightly different meanings. For instance, acquiring rule knowledge may be considered only as one part of *knowledge acquisition*, because the latter may also include other aspects such as acquisition of content knowledge. Nevertheless, the labels can be treated as synonyms within this thesis, because *knowledge acquisition* in problem solving research using LSE-tasks such as MicroDYN is commonly assessed by measuring acquisition of rules (Bühner et al., 2008; Greiff et al., 2012; Kluge, 2008; Kröner et al., 2005).

1.5.2 Preview of Paper 2

Can the construct CPS be measured equally well across Hungarian high school students attending different grades and how does performance differ across these grades? Besides dealing with these questions, this paper investigates incremental validity of MicroDYN in predicting school grades beyond reasoning, expanding research on construct validity in Paper 1 to a sample of high school students. Furthermore, the influence of parental education on performance in CPS is investigated.

As outlined before, evaluating structural stability is a prerequisite for comparing mean differences across groups, a common practice for evaluating psychometric properties of measurement devices assessing intelligence (Chen, 2008), but never applied to measurement devices of CPS. It is expected that students in higher grades outperform students in lower grades. Although sounding trivial, this kind of analyses gives important information on how

[2] The use of the labels rule identification, rule knowledge, and rule application solely applies to Paper 1. In Paper 2 to 4, the CPS dimensions are labelled *use of strategies*, *knowledge acquisition*, and *knowledge application*.

CPS dimensions develop across age, though acknowledging that longitudinal studies have to be conducted to prove and consolidate findings.

Regarding construct validity, it is expected that parental education affects performance in CPS (cf. section 1.4.3). Furthermore, a confirmation of incremental validity of CPS in a sample of high school students would be another indicator of the usefulness of CPS as a predictor of important real world performance outcomes.

1.5.3 Preview of Paper 3

This paper expands research of the second work by comparing results on MicroDYN across samples varying considerably in age as well as in educational background. Does the structure of CPS change across subsamples of junior high school students, senior high school students, university students, and blue-collar workers? And how do their performances differ? CPS tests have never been applied in such a cross-sectional study, making this research rather explorative. Nevertheless, it is expected that senior high school students and university students outperform students in junior high school and blue-collar workers, because performance in CPS is associated with cognitive performance in general. However, it is open to question if work and life experience of older people may compensate for shortcomings in cognitive performance in CPS tasks and how effective this compensation may be.

1.5.4 Preview of Paper 4

The last contribution investigates structural stability and measures performance differences in CPS across gender and nationality by comparing Hungarian and German high school students. Are females or males, and accordingly, Germans or Hungarians, considerably disadvantaged in specific MicroDYN tasks? How does performance in CPS differ across gender and nationality? And if performance differences exist, what are their reasons?

Previous research on CPS differences across gender or nationality usually focussed on analysing mean differences in performance (e.g., Wittmann & Hattrup, 2004) or in *use of strategies* (e.g., Strohschneider & Güss, 1999). However, equality of the measurement device across groups has never been tested before. Thus, it is unclear if differences in performance are based on flaws in the measurement device or real differences in the underlying construct. This paper combines psychometrical analyses of measurement equality across groups with examination of latent mean differences in CPS performance, allowing a sound evaluation of causes of potential differences in performance. Furthermore, the role of *use of strategies* as determinant of performance in MicroDYN is discussed and an outlook for future developments of MicroDYN as well as of other approaches to measure CPS is given.

References

Alexander, K. L., Entwisle, D. R., & Bedinger, S. D. (1994). When expectations work: Race and socioeconomic differences in school performance. *Social Psychology Quarterly, 57,* 283–299.

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

Berry, D. C., & Broadbent, D. E. (1995). Implicit learning in the control of complex systems. In P.A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 131-150). Hillsdale, NJ: Lawrence Erlbaum Associates.

Blech, C., & Funke, J. (2010). You cannot have your cake and eat it, too: How induced goal conflicts affect complex problem solving. *Open Psychology Journal, 3,* 42-53.

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

Byrne, B. M. & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*(2), 287-321.

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.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). Amsterdam, NL: Pergamon.

Chen, H. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.

Chen H., Keith T.Z., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order wisc-iv structure across China, Hongkong, Macau, and Taiwan. *Personality and Individual Differences, 49,* 677–682.

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.

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.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology, 19*(2), 294-304.

Dörner, D. (1986). Diagnostik der operativen Intelligenz [On the diagnostics of operative intelligence]. *Diagnostica, 32*, 290-308.

Fischer, A., Greiff, S., & Funke, J. (2012). The Process of Solving Complex Problems. *Journal of Problem Solving, 4*(1), 19-41.

Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica, 29,* 283–302.

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

Funke, J. (1993). Microworlds based on linear equation systems: A new approach to complex problem solving and experimental results. In G. Strube & K.-F. Wender (Eds.), *The cognitive psychology of knowledge: The German Wissenspsychologie project* (pp. 313-330). Amsterdam: Elsevier.

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

Funke, J. (2003). *Problemlösendes Denken*. Stuttgart: Kohlhammer.

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

Funke, J. (2012). Complex problem solving. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 682-685). Heidelberg: Springer.

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective - 10 years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25-47). New York: Lawrence Erlbaum.

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.

Greiff, S. (2012). *Individualdiagnostik der Problemlösefähigkeit*. [Diagnostics of problem solving ability on an individual level]. Münster: Waxmann.

Greiff, S., Holt, D. V., Wüstenberg, S., Goldhammer, F., & Funke, J. (submitted). Computer-based assessment of Complex Problem Solving: Concept, Implementation, and Application. *Educational Technology and Development.*

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement*, *36*(3), 189-213.

Güss, C. D., Tuason, M. T., & Gerhard, C. (2010). Cross-national comparisons of complex problem-solving strategies in two microworlds. *Cognitive Science, 34,* 489-520.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30,* 304-337.

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

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.

Mayer, R. E. (1992). *Thinking, problem Solving, cognition* (2nd ed.). New Yourk: Freeman.

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.

McGehee, J. J., & Griffith, L. K. (2001). Large-Scale Asesssments combined with curriculum alignment: Agents of Change. *Theory Into Practice, 40*(2), 137-144.

Myrberg, E. & Rosen, E. (2008). A path model with mediating factors of partens' education on students' reading achievement in seven countries. *Educational Research and Evaluation, 14*(6), 507-520.

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

OECD (2009). *PISA 2009 results: What students know and can do: Student performance in Reading, Mathematics, and Science (Volume I).* Paris: OECD.

OECD (2010). *PISA 2012 Problem Solving Framework.* Paris: OECD Publishing.

OECD (2012). *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills.* Paris: OECD.

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

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

Rindermann, H., Flores-Mendoza, C., & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents'

socioeconomic status and education. *Learning and Individual Differences, 20*(5), 544-548.

Rollett, W. (2008). *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme* [Use of strategy, acquisition and application of information, and control of complex dynamic systems]. Münster: LIT.

Sager, S., Barth, C., Diedam, H., Engelhart, M., & Funke, J. (2011). Optimization as an analysis tool for human complex decision making. *SIAM Journal on Optimization, 21,* 936-959.

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.

Schoppek, W. (2004). Teaching structural knowledge in the control of dynamic systems: Direction of causality makes a difference. In K. D. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1219-1224). Mahwah, NJ: Lawrence Erlbaum Associates.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling, 54,* 54-72.

Strohschneider, S. & Güss, D. (1999). The Fate of the Moros: A Cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology, 34*(4), 235-252.

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.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257-285.

Vollmeyer, R. & Burns, B. D. (1996). Hypotheseninduktion und Zielspezifität: Bedingungen, Erlernen und Kontrollieren eines komplexen Systems beeinflussen [Induction of hypothesis and goal specificity: Constraints for learning and controlling complex systems]. *Zeitschrift für Experimentelle Psychologie, 43*(4), 657-683.

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.

Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios - Taxonomie, Entwicklung, Evaluation.* Lengerich: Pabst Science Publishers.

Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55-72). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Wittmann, W. & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393-440.

2

Complex Problem Solving −

More than Reasoning?

Written copyright permission for including this article in the published version of this dissertation was granted November 09, 2012. This article is available as:

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex Problem Solving − More than reasoning? *Intelligence, 40*, 1-14.

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.

© 2011 Elsevier Inc. All rights reserved.

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:

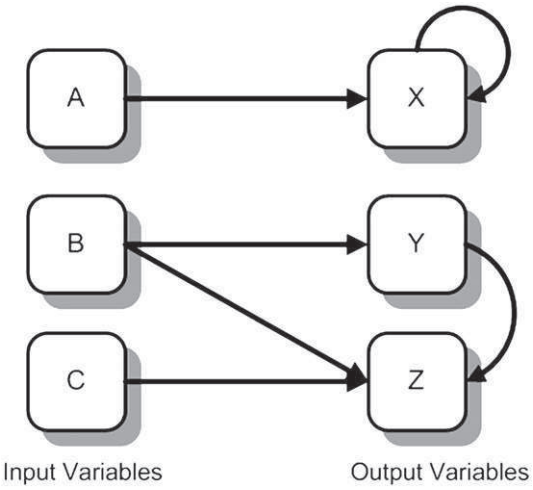

Fig. 1. Structure of a typical MicroDYN item displaying 3 input (A, B, C) and 3 output (X, Y, Z) variables.

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

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

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

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,

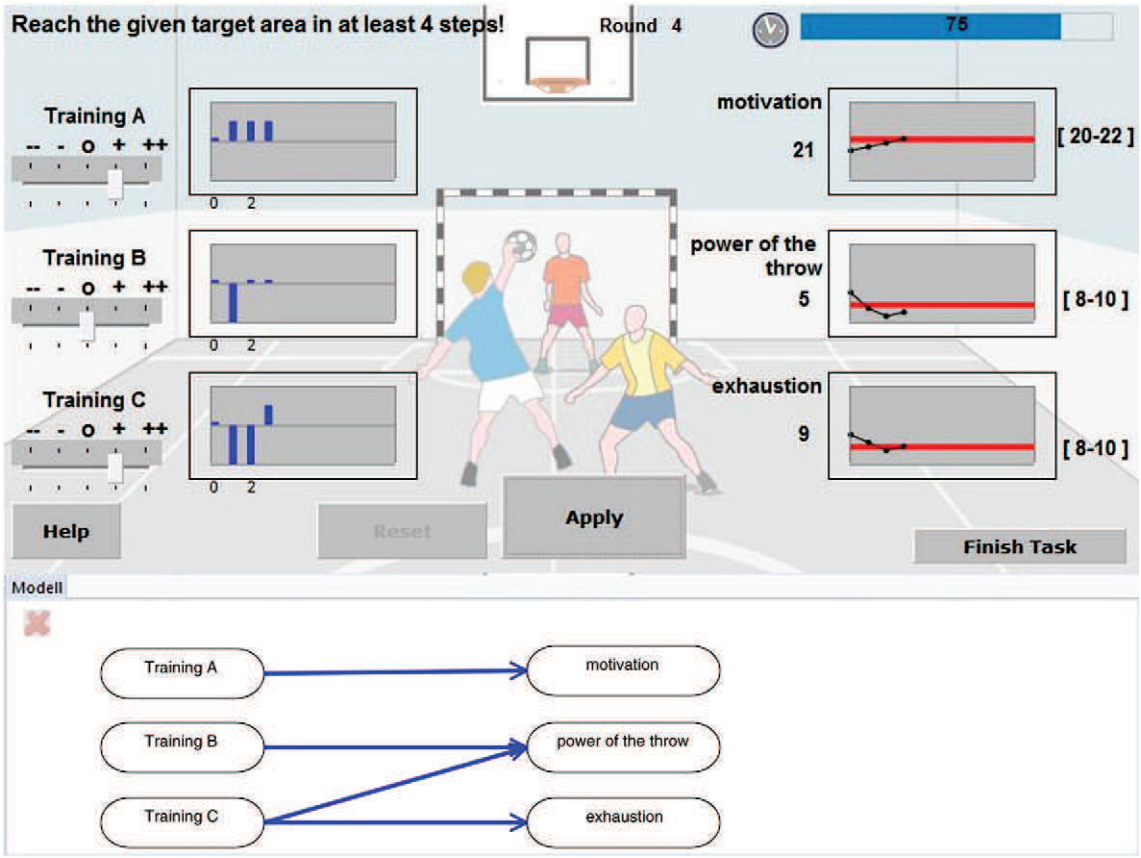

Fig. 2. Screenshot of the MicroDYN-item "handball training" control phase. The controllers of the input variables range from "- -" (value = −2) to "++" (value = +2). The current value is displayed numerically and the target values of the output variables are displayed graphically and numerically.

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

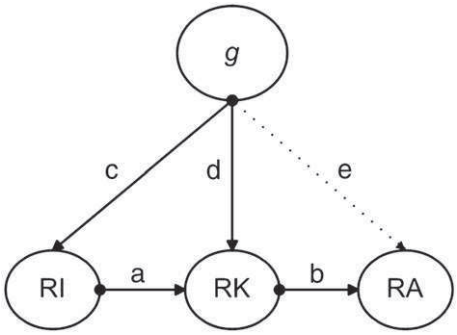

Fig. 3. Theoretical model of the relations between reasoning (g) and the CPS facets rule identification (RI), rule knowledge (RK) and rule application (RA). The dotted line indicates a insignificant path coefficient (e). All four other paths are expected to be significant.

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.

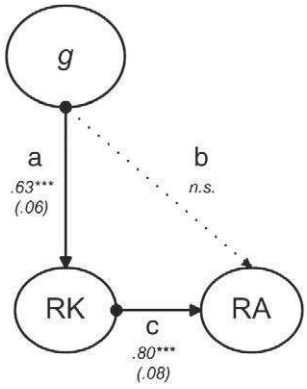

Fig. 4. Structural model including reasoning (g), MicroDYN rule knowledge (RK) and MicroDYN rule application (RA) (n = 222). Manifest variables are not depicted. *p<0.05; **p<0.01; ***p<0.001.

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.

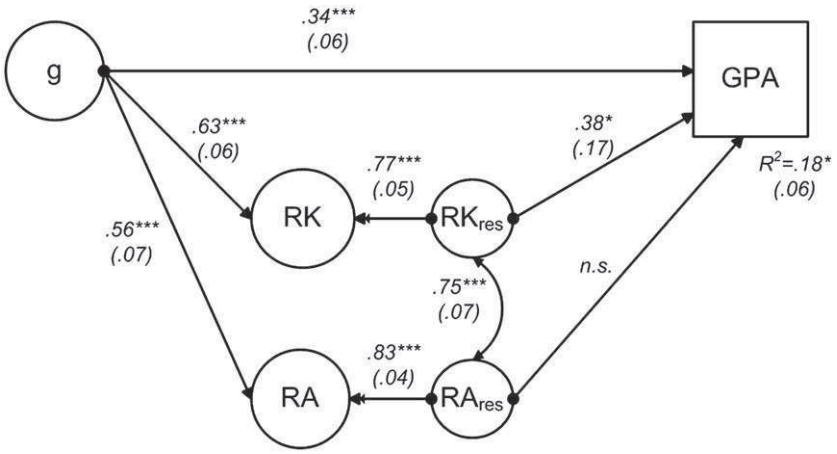

Fig. 5. Rule knowledge (RK) and rule application (RA) were regressed on reasoning (g). The residuals of this regression as well as reasoning were used to predict GPA. Manifest variables are not depicted. *p<0.05; **p<0.01; ***p<0.001.

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.

3

Complex Problem Solving in Educational Settings – Something Beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity.

This article is available as:

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (in press). Complex Problem Solving in Educational Settings – something beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity. *Journal of Educational Psychology.*

Abstract

Innovative assessments of cross-curricular competencies such as complex problem solving (CPS) have currently received considerable attention in large-scale educational studies. This study investigates the nature of CPS by applying a state-of-the-art approach to assess CPS in high school. We analyzed whether two processes derived from cognitive psychology, *knowledge acquisition* and *knowledge application*, could be measured equally well across grades and how these processes differed between grades. Further, relations between CPS, general mental ability (g), academic achievement, and parental education were explored. N = 855 Hungarian high school students in Grades 5 to 11 completed MicroDYN, which is a computer-based CPS test, and the Culture Fair Test 20-R as a measure of g. Results based on structural equation models showed that empirical modeling of CPS was in line with theories from cognitive psychology such that the two dimensions identified above were found in all grades, and that there was some development of CPS in school although the Grade-9 students deviated from the general pattern of development. Finally, path analysis showed that CPS was a relevant predictor of academic achievement over and above g. Overall, results of the current study provide support for an understanding of CPS as a cross-curricular skill that is accessible through computer-based assessment and that yields substantial relations to school performance. Thus, the increasing attention CPS has currently received on an international level seems warranted given its high relevance for educational psychologists.

Complex Problem Solving in Educational Contexts—Something Beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity

Improving students' minds is considered a major challenge in education. One way to achieve this is by enhancing students' problem-solving skills (Mayer & Wittrock, 2006), which are captured in their ability to solve novel problems. The importance of problem solving for success in life is also reflected in the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development (OECD), which is allegedly the most comprehensive and important international large-scale assessment in existence today (e.g., OECD, 2004; 2010). The PISA studies aim to evaluate educational systems worldwide by assessing 15-year-olds' competencies in the key subjects reading, mathematics, and science, and also to evaluate more complex cross-curricular skills such as complex problem solving (e.g., OECD, 2010).

Specifically, cross-curricular complex problem solving (CPS) was assessed in more than half a million students in over 70 countries (e.g., OECD, 2009) in the current PISA 2012 cycle.[1] As an example of a typical CPS task in PISA 2012, imagine that you just bought your first mobile phone ever, you have never worked with such a device, and now you want to send a text message. Essentially, there are two things you need to do: (a) press buttons in order to navigate through menus and to get feedback on your actions and (b) apply this knowledge to reach your goal, that is, to send a text message. These aspects of CPS are also reflected in Buchner's (1995) definition:

> Complex Problem Solving 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

[1] In PISA, the term *interactive problem solving* (OECD, 2010) is used. Other labels referring to the same construct are *dynamic problem solving,* which focuses on the aspect of systems to change dynamically (e.g., Greiff, Wüstenberg, & Funke, 2012) and *complex problem solving* (Dörner, 1986, 1990), which emphasizes the aspect of the underlying system's complexity. In the present paper, we use the term *complex problem solving (CPS)*, which is most established in research.

> regularities can only be revealed by successful exploration and integration of the information gained in that process. (p. 14)

Funke (2010) and Raven (2000) concluded that CPS requires a series of complex cognitive operations such as planning and implementing actions, model building, or self-regulation. Enhancing these cognitive operations is the goal of any educational system, or, as Mayer and Wittrock (2006) put it: "One of educational psychology's greatest challenges [is to help] students become better problem solvers" (p. 299). However, CPS research that combines assessment and theory is rather scarce. The present study contributes to research on the nature and validity of CPS by applying a state-of-the-art approach to assess CPS in high school students.

Complex Problem Solving and g

Research on general mental ability, now often referred to as psychometric g, was also initially educationally motivated. That is, when Alfred Binet and Théodore Simon (1905) developed the first psychometric tests of g, their starting point was to objectively identify students with learning disabilities who were in need of specially tailored education. Ever since then, no other construct has been as extensively and continuously validated in educational contexts. Specifically, based on the assorted existing empirical evidence, Reeve and Hakel (2002) concluded that there is a common mechanism underlying human mental processes labelled psychometric g. Only a few researchers have recently challenged this view by questioning the importance of g or by introducing alternative concepts such as practical intelligence (e.g., Lievens & Chan, 2010), social intelligence (e.g., Kihlstrom & Cantor, 2011), or emotional intelligence (e.g., Goleman, 1995). That is, the overwhelming conceptual and empirical evidence has supported the educational importance of g concerning manifold external criteria. The most impressive accumulation of evidence was provided by Ree and Carretta (2002) who related skills, personality, creativity, health, occupational status, and income to measures of g.

Theoretically, g is bolstered by the Cattell-Horn-Carroll (CHC) theory, which assumes that g is on a general level of cognitive ability (Stratum III), which in turn influences about 10 broad cognitive abilities on the second level (Stratum II). Narrow cognitive abilities are located on the lowest level (Stratum I; McGrew, 2009). CHC theory is considered particularly relevant to school psychologists and other practitioners for educational assessment and has received considerable attention in the educational arena. On a measurement level, strict requirements such as structural stability have been frequently shown to hold for tests of g (e.g., Taub & McGrew, 2004). Structural stability indicates that the construct does not change across groups and that test scores do not depend on the group to which the test is administered (Byrne & Stewart, 2006). This is a prerequisite for interpreting differences in mean performance (Cheung & Rensvold, 2002). In light of the overall empirical and theoretical evidence, it is not surprising that Reeve and Hakel (2002) consider g to be crucial in any educational context.

However, the predominant role of g in education has not been entirely undisputed. Whereas Sternberg (1984, 2009) proposed a triarchic theory of intelligence composed of an analytical, a practical, and a creative component, Diaz and Heining-Boynton (1995) noted the relevance of alternative concepts such as CPS for students' education; thus, going beyond the idea that a single mental construct underlies cognitive performance and including more complex processes. The general rationale behind this idea is that despite the well-established predictive power of g, many questions about its nature remain unsolved (e.g., genetic endowment, environmental influence, different forms of intelligence; Neisser et al., 1996). In fact, g's ability to predict nonacademic performance is considerable but far from perfect even after controlling for measurement error; thus, variance that may be accounted for by CPS is left unexplained (e.g., Rigas, Carling, & Brehmer, 2002). In this context, some studies on the relation between measures of CPS and g have yielded low relations. For example, Putz-Osterloh (1981) reported zero correlations between performance in the CPS scenario

Tailorshop and the *Advanced Progressive Matrices* (Raven, 1962). Even though methodological issues might have caused this result, current findings have supported the distinction between g and CPS and have demonstrated the added value of CPS beyond g in different contexts (e.g., Danner, Hagemann, Schankin, Hager, & Funke, 2011; Wüstenberg, Greiff, & Funke, 2012).

Even during the early stages of CPS research, Dörner (1986) criticized the focus on the speed and accuracy of the capacity for basic information processing in measures of g (e.g., Raven, 2000) and suggested that a stronger emphasis be placed on the strategic and processing aspects of the mental processes involved in CPS. He proposed measuring complex cognitive processes in CPS to overcome the "out-of-touch-with-reality" issue that traditional intelligence tests suffer from (Dörner & Kreuzig, 1983). The broad conception of mental ability in CPS connects directly to the understanding of learning in the classroom. Mayer and Wittrock (2006) stated that a deep understanding of the nature of problem solving is needed if meaningful learning is to be fostered. Thus, going beyond current conceptualizations of g, meaningful learning and problem solving are closely related (Sternberg, 2000), and they are of great importance to both predict and understand complex learning processes in classrooms (Mayer & Wittrock, 2006). Similar to Sternberg and his conception of intelligence (1984, 2009), the line of research on CPS that emerged around Dörner (1986) does not seriously object to the use of measures of g, but suggests complementing them with additional measures such as CPS and its defining cognitive processes.

Complex Problem Solving in Cognitive Science

Mayer (2003) defined problem solving in general as transforming a given state into a goal state when no obvious method of solution is available. According to Funke and Frensch (2007), a problem solver has to overcome barriers by applying operators and tools to solve a problem. However, problem solving may take place in different educationally relevant domains, and a large body of research has been conducted in domain-specific areas such as

mathematical, scientific, or technical problem solving (Sugrue, 1995). Besides these domain-specific approaches, the idea of domain-general processes generally involved in problem solving was taken up by the European line of research on complex problem solving mentioned above (e.g., Dörner, 1986; Funke, 2001; Funke & Frensch, 2007).

This line of research assumes that domain-general processes are crucial when participants deal with an unknown and highly interrelated system (i.e., a complex problem) for the first time, although when dealing with the same problem repeatedly, domain-specific knowledge may be increasingly involved. That is, CPS research acknowledges that previous experiences or the problem context may influence CPS, but these aspects are not of elementary concern, and problems are designed to be solvable without prior knowledge. In the tradition of Newell and Simon (1972), who described problem-solving behavior uncontaminated by domain-specific knowledge, CPS research aims to uncover general cognitive processes before a considerable amount of prior knowledge is gathered and, thus, before problem solvers switch to more specialized strategies.

Generally, two main demands specify a problem solver's performance within the realm of CPS: knowledge acquisition and knowledge application (Funke, 2001). For instance, dealing with an entirely new mobile phone as outlined previously describes a specific situation that is typically considered to be a complex problem involving dynamic interaction with a yet-unknown system in order to (a) acquire knowledge and (b) use this knowledge for one's own purposes. This delineation into two main cognitive processes is not only logical and has been widely applied when assessing CPS (e.g., Fischer, Greiff, & Funke, 2012; Funke, 2001; Kröner, Plass, & Leutner 2005), but it also connects to general research on (a) problem representation and (b) the generation of problem solutions.

Regarding problem representation, the Gestalt psychologist Duncker (1945) was the first to emphasize the importance of a sound problem representation, and Markman (1999) has further elaborated on this concept. According to Markman's elaboration, a representation

begins with a description of the elements of a complex problem, the *represented world*, and a set of operators that can be used to relate these elements to each other, the *representing world.* Represented and representing worlds are usually predefined in CPS research, that is, the problems are well defined (represented world), and the set of operators available is limited and can be used only within given constraints (representing world; this setup is often found in educational contexts; Mayer & Wittrock, 2006). The elements of a complex problem (represented world) and the set of operators (representing world) are subsequently connected by a set of rules that are established while the problem solver attempts to penetrate the problem. This kind of task is often required of students in school and is at the core of the solver's task in CPS. It describes the process of building a problem representation. In the example above, a description of the problem (i.e., sending a text message) and the set of elements (i.e., inputs and outputs of the mobile phone) are predefined, but the connections between them are yet to be built. Finally, this needs to lead into a process that uses the representation that was established before the problem solution (Markman, 1999). It is this representational function that gives meaning to the representation (Novick & Bassok, 2005) and that constitutes the link between the problem representation (i.e., knowledge acquisition) and generating a problem solution (i.e., knowledge application).

Regarding the generation of a problem solution, algorithmic and heuristic strategies represent a common distinction between different types of solutions. Whereas algorithms are guaranteed to yield a solution, heuristics are usually applied when an exhaustive check of all possible moves is not efficient (Novick & Bassok, 2005). As this exhaustive check is scarcely possible in complex problems, it is safe to assume that the process of solving them is largely guided by heuristics such as a mean-ends analysis (Newell & Simon, 1972). In fact, Greeno and Simon (1988) stated that problem solvers tend to prefer a mean-ends analysis as the solution method when faced with novel problems that are relatively free of prior knowledge and in which well-defined goals are given. Often, when students face transfer problems in

educational contexts, it is under exactly the condition that prior factual knowledge is of limited help in solving the problem at hand and that the available operators are clearly defined (Mayer & Wittrock, 2006).

Obviously, knowledge acquisition and knowledge application are closely entangled because a good representation is to a certain degree a necessary condition for establishing specific goals and for deducing interventions to solve a problem (Novick & Bassok, 2005). Thus, researchers in both of the two aforementioned fields have emphasized the importance of the respective aspect: Newell and Simon (1972) introduced the concept of a *problem space* in which the problem, its rules, and its states are represented, focusing on aspects of knowledge acquisition. By contrast, Markman (1999) considered the use of information essential and, thus, the process of knowledge application. Novick and Bassok (2005) stated that "although it is possible to focus one's research on one or the other of these components, a full understanding of problem solving requires an integration of the two" (p. 344). As it is widely acknowledged that representation and solutions interact with each other, the neglect of concrete efforts to converge these two lines of research has been surprising.

Measurement Approaches to Complex Problem Solving

A comprehensive assessment of the CPS dimension knowledge acquisition requires the active exploration of an unknown system, and assessment of knowledge application requires the immediate adaption to actions initiated by the system. Thus, by definition, the assessment of CPS is always computer-based as the task changes interactively by itself or due to the user's intervention (Funke & Frensch, 2007), which cannot be assessed on a paper-pencil basis (Funke, 2001).

Consequently, computer-based microworlds (e.g., Gardner & Berry, 1995) were developed to reliably measure CPS performance. However, most efforts were overshadowed by severe measurement issues (cf. Greiff, Wüstenberg, & Funke, 2012; Kröner et al., 2005). It was only recently that multiple complex systems were introduced as another advance in the

assessment of CPS (Greiff et al., 2012). In a multiple complex systems approach, time on each task is significantly reduced and tasks are directly scaled with regard to their difficulty (Greiff, 2012). Hence, in one testing session, problem solvers work on several independent tasks and are confronted with an entire battery of CPS tasks. In this manner, a wide range of tasks with varying difficulty can be employed, leading to increased reliability. Thus, the theoretically derived internal structure of CPS with its distinction between knowledge acquisition and knowledge application was able to be psychometrically confirmed for the first time with the advent of multiple complex systems (e.g., Greiff et al., 2012). The difference between measures of g and CPS in terms of discriminant and predictive validity could also be accounted for (Sonnleitner et al., 2012; Wüstenberg et al., 2012).

Multiple Complex Systems within the MicroDYN Approach

MicroDYN is an example of a test battery that is based on multiple complex systems within the linear structural equation (LSE) framework (Funke, 2001). In LSE tasks, the relations between input variables and output variables are described by linear structural equations. However, in MicroDYN, time per task is considerably shorter than for classical LSE tasks (Funke, 2001), thus allowing for a sufficient number of problems to be attended to in order to achieve acceptable measurement. Problem solvers face seven to nine tasks, each lasting about a maximum of 5 min, which sums to an overall testing time of approximately 45 min including instruction. MicroDYN tasks consist of up to three input variables (denoted by A, B, and C), which are related to up to three output variables (denoted by X, Y, and Z; see Figure 1), but only the former can be directly manipulated by the problem solver (Greiff, 2012; Wüstenberg et al., 2012).

Please insert Figure 1 about here

Input and output variables can be related to each other in different ways, however, these relations are not apparent to the problem solver. Causal relations between input variables and output variables are called *direct effects*, whereas effects originating and ending with output

variables are called *indirect effects.* The latter involve side effects (see Figure 1: Y to Z) when output variables influence each other and eigendynamics (see Figure 1: X to X) when output variables influence themselves. Problem solvers cannot influence these two effects directly, however, the effects are detectable through the adequate use of strategy. All tasks have different cover stories, and the names of input and output variables are labelled either fictitiously (e.g., Brekon as a name for a specific cat food) or without deep semantic meaning (e.g., red butterfly as the name of a butterfly species). For instance, in the task *game night* (see Figure 2), different kinds of chips labelled blue, green, or red chip serve as input variables, whereas different kinds of playing cards labelled Royal, Grande, or Nobilis serve as output variables.

Please insert Figure 2 about here

While working on a MicroDYN task, a problem solver faces two different phases. In Phase 1, problem solvers can freely explore the system by entering values for the input variables (e.g., varying the amount of blue, green, and red chips in Figure 2). This is considered an evaluation-free exploration, which allows problem solvers to engage with the system and to use their knowledge acquisition ability under standardized conditions without controlling the system (Kröner et al., 2005). During this Phase 1, problem solvers are asked to draw the connections between variables onscreen (see bottom of Figure 2), thereby producing data reflecting the knowledge acquired (3 min for Phase 1). Mayer (2003) calls this a situational external representation of a problem. In this first phase, the amount and correctness of explicit knowledge gathered during exploration is measured and expressed in a mental model as the final external problem representation (Funke, 2001). In Phase 2, problem solvers are asked to reach given target values on the output variables (e.g., card piles Royal, Grande, and Nobilis in Figure 2) by entering correct values for the input variables, thereby producing data reflecting the application of their knowledge (1.5 min for Phase 2). In this second phase, the goal-oriented use of knowledge is assessed.

These two phases are directly linked to the concepts of knowledge acquisition (i.e., representation) and knowledge application (i.e., generating and acting out a solution; Novick & Bassok, 2005). More detailed information on both the underlying formalism and the MicroDYN approach can be found in Funke (2001), Greiff et al. (2012), and Wüstenberg et al. (2012).

Multiple complex systems as implemented in MicroDYN were used internationally to assess the CPS ability of 15-year-old students in the 2012 cycle of PISA. Clearly, the necessary steps toward computer-based assessment in large-scale assessments comes along with great potential (Kyllonen, 2009), yet many questions about the nature of CPS and its measurement characteristics remain unanswered.

Purpose of Study and Hypotheses

The present article is aimed at advancing knowledge of CPS, its assessment, and its use in educational contexts. Specifically, the accuracy, precision, and usefulness of test scores derived for educational purposes depend on theoretical support and good psychometric properties (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Benson, Hulac, & Kranzler, 2010). That is, if one wants to adequately interpret students' CPS scores, a sound assessment device is needed. This has not been sufficiently established for CPS and is just beginning to emerge in the form of multiple complex systems. The purpose of this study was fourfold and was aimed at elaborating the construct of CPS and its operationalization as defined above in a representative sample of Hungarian students. Specifically, we tested (1) the underlying dimensionality, assuming a measurement model with two different CPS processes (i.e., knowledge application and knowledge acquisition), (2) the structural stability of the CPS construct across different grade levels of high school students aged 11 to 17, (3) latent mean comparisons between these grade levels if measurement invariance was sufficiently met, and

(4) structural relations between CPS, fluid intelligence as a proximal measure of g, GPA, and parental education across these groups to assess construct validity.

(1) With regard to the dimensionality of CPS, a large body of conceptual literature has suggested that the two CPS processes, knowledge acquisition and knowledge application, are related and yet somewhat distinct aspects of an overarching CPS process, but empirically this has been shown only for very selective samples (e.g., Kröner et al., 2005; Wüstenberg et al., 2012) and not yet for high school students. As part of assessing structural validity, we adhered to the question of whether there is an adequate construct representation of CPS by testing a measurement model that was closely aligned with the idea of partially separate mechanisms for problem representation and problem solution and assumed a 2-dimensional model composed of the dimensions knowledge acquisition and knowledge application.

Hypothesis (1): We expected CPS to be composed of two different processes, knowledge application and knowledge acquisition. Thus, a 2-dimensional model was expected to show the best fit and to fit significantly better than a 1-dimensional model with the two processes combined under one first-order factor.

(2) The structural stability of CPS pertains to the exact nature of the construct assessed. That is, the structure of the construct was not expected to change across different grade levels, indicating that the interpretation of test scores does not depend on the specific group the test is administered to (Byrne & Stewart, 2006). This was tested by evaluating measurement invariance. Only to the extent that measurement invariance exists, between-group differences of grade levels are unambiguous and can be interpreted as true and not as psychometric differences in latent ability (Cheung & Rensvold, 2002, cf. Results section). For instance, it may be that due to cognitive development that occurs during adolescence, the construct of CPS changes. Analyses of measurement invariance would show that tasks behave differently in groups in different grade levels just as self-ratings on questionnaires may change their meaning when questions are translated from one culture to another (Chen, 2008).

Hypothesis (2): We expected CPS to show measurement invariance across different school grades.

The aspect of measurement invariance led directly to (3) latent mean comparisons of different grade levels or, in other words, to the question of level stability. For those parts of the measurement model that are identified as invariant, latent factor means can be compared, thus providing important insights into the effects of schooling and environment on CPS.

Hypothesis (3): If measurement invariance was sufficiently met, we expected latent mean differences between groups to indicate that students in higher grades perform significantly better in knowledge acquisition and knowledge application than students in lower grades.

In addition to establishing the validity of the internal structure, another important step is (4) establishing construct validity in terms of divergent and convergent relations to other constructs. To this end, we assessed how CPS was related to a measure of g, GPA, and parental education. Whereas GPA is an excellent marker of academic achievement, parental education reflects one of the most important socioeconomic variables with a strong impact on school performance and educational outcomes (Myrberg & Rosen, 2008).

Hypothesis (4): Concerning construct validity, we expected that (a) g would predict performance on CPS tasks. However, a considerable amount of CPS variance was expected to remain unexplained suggesting that parts of CPS are independent from g and that (b) CPS predicted GPA beyond g as indicated by conceptual considerations and previous research. Furthermore, we expected that (c) parental education would predict performance in CPS and in g.

The field of CPS lags behind the field of intelligence testing, in which a broad range of well established and extensively validated assessment procedures exist, some of which are even specifically tailored to educational demands (e.g., Wechsler, 2008). Considering the current educational interest in the assessment of CPS and the associated implications for

researchers as well as practitioners such as educators and policy makers, this is particularly troublesome. By addressing the four research questions above, we aimed to make the measurement of CPS more evidence-based, thereby helping the field of CPS to catch up.

Method

Participants

Our sample ($N = 855$) was a subsample of a larger and representative study ($N > 4{,}000$) conducted in Hungary. Participants were randomly drawn from Grades 5 to 11 in Hungarian elementary (Grades 5 to 8) and secondary schools (Grades 9 to 12).

Some software problems occurred during online testing, resulting in data loss. However, data were missing completely at random. Participants who were missing more than 50% of their data on MicroDYN or any other measure were excluded from all analyses (only about 5% of participants provided less than 80% data); other missing data were excluded on a pairwise basis.

Finally, data from 855 students were available for the analyses of Hypotheses 1 to 3 (age $M = 14.11$; $SD = 1.83$; 46% male). However, all analyses including g (Hypotheses 4a and 4b) were based on a smaller subsample of students who completed both tests of CPS and g ($N = 486$; age: $M = 14.36$; $SD = 1.16$; 45% male). Data were missing by design because g was not assessed in Grades 5, 6, and 11, and only a small number of missing values occurred due to drop-out (e.g., illness of students).

Design

CPS. MicroDYN was administered on computers. At the beginning, participants were instructed how to complete a trial task, in which they learned how the interface of the program could be controlled and which two tasks they were expected to solve: Participants explored unknown systems and drew their conclusions about how variables were interconnected in a situational model (cf. bottom of Figure 2; Mayer, 2003). This situational model was seen as an appropriate way of representing gathered information and allowed

participants to visualize their mental model (knowledge acquisition; Funke, 2001). Subsequently, they controlled the system by reaching given target values (knowledge application). After having finished the instruction phase, participants were given eight consecutive MicroDYN tasks. One task had to be excluded from analyses due to low communality (r^2 = .03) caused by an extreme item difficulty on knowledge acquisition (*P* = .03). All subsequent analyses were based on seven tasks. The task characteristics of all tasks (e.g., number of effects) were varied to produce tasks with an appropriate difficulty for high school students (cf. Greiff et al., 2012; see Appendix for equations).

g. The *Culture Fair Test 20-R* (CFT) consists of four subscales that measure fluid intelligence, which is seen as an excellent marker of g (Weiß, 2006) and is assumed to be at the core of intelligence (Carroll, 2003).

Dependent Variables and Scoring

CPS. Both MicroDYN dimensions, knowledge acquisition and knowledge application, were scored dichotomously, which is an appropriate way to score CPS performance (see Greiff, et al., 2012; Kröner et al., 2005; Wüstenberg et al., 2012). For knowledge application, users' models were evaluated and credit was given for a completely correct model, whereas no credit was given when a model contained at least one mistake. Knowledge application was scored as correct when all target values of the output variables were reached.

g. All items of the CFT were scored dichotomously according to the recommendations in the manual (Weiß, 2006).

GPA and Parental Education. Participants self-reported their GPA from the previous school year and the educational levels of their parents. GPA ranged from 1 (*insufficient*) to 5 (*best performance*). Parental educational level for both mothers and fathers was scored on an ordinal scale (1 = no elementary school graduation; 2 = elementary school; 3 = secondary school; 4 = university-entrance diploma; 5 = lower level university; 6 = normal university; 7 = PhD).

Procedure

Test execution took place in the computer rooms of the participating Hungarian schools and lasted approximately 90 min. Participants worked on MicroDYN first, and the CFT was administered afterwards. Finally, participants provided demographic information. MicroDYN was delivered through the online platform *Testing Assisté par Ordinateur* (computer-based testing). Testing sessions were supervised either by research assistants or by teachers who had been trained in test administration.

Results

Descriptive Statistics

Analyses of manifest variables showed that the internal consistencies of MicroDYN as measures of CPS were acceptable (knowledge acquisition: $\alpha = .75$; knowledge application: $\alpha = .74$) and Cronbach's α for the CFT ($\alpha = .88$) was good. Participants' raw score distributions on the CFT ($M_7 = 39.84$, $SD = 9.13$; $M_8 = 41.36$, $SD = 7.54$; $M_9 = 36.97$, $SD = 7.20$; $M_{10} = 38.37$, $SD = 8.02$) differed slightly compared to the original scaling sample of students attending the same grades ($M_7 = 34.98$, $SD = 6.63$; $M_8 = 36.37$, $SD = 6.56$; $M_9 = 38.42$, $SD = 6.43$; $M_{10} = 39.31$, $SD = 6.90$; Weiß, 2006). Further, participants' GPA showed a sufficient range ($M_7 = 4.00$, $SD = 0.80$; $M_8 = 3.95$, $SD = 0.83$; $M_9 = 3.64$, $SD = 1.05$; $M_{10} = 3.77$, $SD = 0.74$; $M_{11} = 3.64$, $SD = 0.71$; 1 = insufficient, 5 = best performance), and so did mothers' and fathers' education scores ($M_{\text{Mother}} = 3.85$, $SD = 1.09$; $M_{\text{Father}} = 3.75$, $SD = 1.10$; 1 = no elementary school graduation, 7 = PhD).

Statistical Analyses and Data Transformation

The analyses on the dimensionality of CPS (Hypothesis 1), measurement invariance (Hypothesis 2), latent mean differences (Hypothesis 3), and construct validity including only CPS and g (Hypothesis 4a) were based on latent models using structural equation modeling (SEM; Bollen, 1989). SEM analyses using latent variables require larger sample sizes than traditional statistics based on manifest variables. On this matter, Ullman (2007) recommended

that the number of estimated parameters should be no more than one fifth of *N*. To meet this guideline, we merged Grades 5 and 6, 7 and 8, as well as 10 and 11, to *grade levels 5/6, 7/8, and 10/11*, so that sufficient data were provided to test measurement models separately within each group or grade level, respectively. We kept Grade 9 as a single grade level because the transition from elementary to secondary school takes place after Grade 8 in the Hungarian school system. This transition is known to affect cognitive performance and to be associated with a general loss in achievement (e.g., Alspaugh & Harting, 1995; Smith, 2006). Specifically, Molnár and Csapó (2007) reported a drop in problem solving performance in Grade 9 test scores in Hungary. Even though we did not pose any hypotheses about the performance pattern in Grade 9, we did not merge these students in order to be able to detect effects of the transition. All other analyses including GPA, CPS, g, and parental education (Hypothesis 4b and 4c) were based on manifest (observed) data (cf. results on Hypotheses 4b and 4c). Mplus 5.0 was used for all analyses (Muthén & Muthén, 2010).

Hypothesis 1: Dimensionality of CPS

We used confirmatory factor analyses within SEM to test the underlying measurement model of CPS with the two different CPS processes knowledge acquisition and knowledge application (Hypothesis 1). Table 1 shows the dimensionality results.

Please insert Table 1 about here

The 2-dimensional model fit well in the overall sample compared to cut-off values recommended by Hu and Bentler (1999), who stated that Comparative Fit Index and Tucker Lewis Index (TLI) values above .95 and a Root Mean Square Error of Approximation (RMSEA) below .06 indicate a good global model fit. Within the 2-dimensional model, the measures of knowledge acquisition and application were significantly correlated on a latent level ($r = .74$, $p < .001$; manifest correlation: $r = .52$, $p > .001$). When estimating this and all subsequent models, we used the preferred estimator for categorical variables: the Weighted Least Squares Mean and Variance adjusted estimator (WLSMV; Muthén & Muthén, 2010).

We also tested a 1-dimensional model with all indicators combined under one general factor; however, the fit indices decreased considerably. In order to compare the 2-dimensional and 1-dimensional model, χ^2 values in Table 1 cannot be directly subtracted to compare them because computing the differences of χ^2 values and *df* between models is not appropriate if WLSMV estimation is applied (Muthén & Muthén, 2010, p. 435). Thus, we carried out a χ^2 difference test in Mplus (Muthén & Muthén, 2010), which showed that the 2-dimensional model fit significantly better than the 1-dimensional model ($\chi^2 = 86.121$; $df = 1$; $p < .001$). After this, the 2-dimensional model was applied to each grade level (i.e., grade levels 5/6, 7/8, 9, and 10/11) separately, also showing a very good fit (Table 1).

In summary, the 2-dimensional model fit well in the overall sample and for each grade level. Thus, the processes knowledge acquisition and knowledge application were empirically distinguished, supporting Hypothesis 1.

Measurement Model of g

As a prerequisite for all analyses involving g, we had to test a measurement model for the CFT. Because the CFT contains 56 items, we decided to use the item-to-construct balance recommended by Little, Cunningham, Shahar, and Widaman (2002) to assign items to four parcels. Each parcel consisted of 14 CFT items to reduce the number of parameters to be estimated. The mean difficulty of the parcels did not differ significantly ($M_1 = .72$; $M_2 = .75$; $M_3 = .71$; $M_4 = .66$; $F_{3, 56} = 0.52$; $p > .05$) and the parcels' factor loadings were also comparable ($\beta_1 = .82$, $\beta_1 = .78$, $\beta_1 = .80$, $\beta_1 = .78$, $F_{3, 56} = 0.33$; $p > .05$). The measurement model with g based on four parcels showed a very good fit for the overall sample ($N = 486$; $\chi^2 = .717$; $df = 2$; $p > .05$; CFI = .999; TLI = .999; RMSEA = .001), as well as for the different grade levels (CFI = .991-.999; TLI = .991-.999; RMSEA = .001-.002).

Hypothesis 2: Measurement Invariance

Measurement invariance was tested by multigroup analyses using the means and covariance structure (MACS) approach within SEM. The general procedure of testing measurement invariance is explained in detail by Byrne and Stewart (2006). They describe a series of hierarchical steps that have to be carried out such that each step imposes an increasingly greater number of restrictions to model parameters to test invariance. Thereby, four different models of invariance are distinguished: configural invariance, weak factorial invariance, strong factorial invariance, and strict factorial invariance. In general, measurement invariance is met if restrictions of model parameters in one model do not generate a substantially worse model fit in comparison to an unrestricted model. The model fit can be evaluated by either a *practical* perspective, reflected in a drop in fit indices such as the CFI (CFI < .01; Cheung & Rensvold, 2002), or by a stricter *traditional* approach, indicated by a significant χ^2 difference test. Only if at least strong factorial invariance is established can latent mean comparisons (Hypothesis 3) be meaningfully interpreted. Otherwise, between-group differences may reflect psychometric properties of the items and not true differences (Byrne & Stewart, 2006).

CPS. To test the measurement invariance of MicroDYN, we applied a procedure that is slightly different from the typical one recommended by Byrne and Stewart (2006). MicroDYN data were based on categorical variables and, thus, constraints on model parameters differed in comparison to invariance tests based on continuous variables (Muthén & Muthén, 2010, p. 433).

Indices of global model fit for all analyses on measurement invariance are shown in Table 2. Based on the 2-dimensional model derived in Hypothesis 1, this multigroup model testing configural invariance of CPS fit well. In this model, thresholds[2] and factor loadings were not constrained across groups, factor means were fixed to zero in all groups, and

[2] In models containing categorical variables, thresholds are used instead of intercepts.

residual variances were fixed to one in all groups (as recommended by Muthén & Muthén, 2010, p. 434) instead of freely estimating residuals as it is done with continuous outcomes.

Please insert Table 2 about here

Weak factorial invariance was not tested because it is not recommended when the WLSMV estimator for categorical outcomes is used (Muthén & Muthén, 2010, p. 433). Thus, the next step was to test for strong factorial invariance, in which thresholds and factor loadings were constrained to be equal across groups, residual variances were fixed to one, and factor means were fixed to zero in one group (i.e., grade level 5/6), whereas there were no constraints specified in any other group (Muthén & Muthén, 2010, p. 343). The strong factorial invariance model did not show a decrease in model fit based on the practical perspective (ΔCFI < .01) or based on the stricter traditional perspective (nonsignificant χ^2 difference test, Table 2) compared to the configural invariance model. Finally, we evaluated strict factorial invariance, in which, in addition to the restrictions realized in strong factorial invariance, all residual variances were fixed to one in all groups. Results from Table 2 showed that MicroDYN was also invariant in a strict sense, even though strict factorial invariance is not a prerequisite for group comparisons of latent factor means and variances (see Byrne & Stewart, 2006).

Although invariance was found for MicroDYN, suggesting an identical factor structure across grade levels, single path coefficients can differ without compromising the invariance of the overall model. This would account for correlations between measures of knowledge acquisition and knowledge application, which varied across the different grade levels ($r_{5/6}$ = .82, *SE* = .05; $r_{7/8}$ = .68, *SE* = .05; r_9 = .94, *SE* = .06; $r_{10/11}$ = .72, *SE* = .05). The two dimensions correlated significantly higher in grade level 9 than in grade level 5/6 (based on *z*-statistics), which in turn showed a significantly higher correlation than grade levels 10/11 and 7/8, whereas the latter two did not differ significantly (10/11 = 7/8 < 5/6 < 9). These findings

raised some concerns about the pattern of results for Grade 9; these will be discussed later on in more detail.

In summary, MicroDYN showed measurement invariance so that latent mean differences could be interpreted as true differences in the construct being measured (Byrne & Stewart, 2006). Consequently, Hypothesis 2 was supported.

g. As a prerequisite for Hypothesis 4, we tested for construct validity between CPS, g, and external criteria. At this stage, we also checked for the measurement invariance of the CFT as described in the Method section (and recommended by Byrne & Stewart, 2006). We used Maximum Likelihood estimation for continuous variables for all models because CFT data were parceled and could be considered continuous. The CFT was invariant in a strict sense as indicated by a nonsignificant χ^2 difference ($p > .10$) between the models of strict factorial invariance ($\chi^2 = 25.546$, $df = 26$; CFI = .999, TLI = .999, RMSEA = .001) and configural invariance ($\chi^2 = 3.908$, $df = 6$; CFI = .999, TLI = .999, RMSEA = .001).

Hypothesis 3: Latent Mean Comparisons

CPS. As a prerequisite for comparing means across groups, the MicroDYN scale had to be fixed to a user-specified level by setting the latent means of a reference group to zero in both dimensions (e.g., grade level 5/6), whereas the latent means of all other groups were freely estimated and subsequently compared to the reference group. Thus, we used the strong factorial invariance model and compared all grade levels with each other, starting with grade level 5/6 as the reference group (left part of Table 3), whereas grade level 7/8 served as the reference group in a second comparison (middle part of Table 3) and grade level 9 in a third comparison.

Please insert Table 3 about here

It was expected that all latent means would have a positive value and would differ significantly from the corresponding reference groups, thereby indicating that students in higher grade levels performed better.

Results for measures of knowledge acquisition indicated that grade level 9 performed worse than grade level 5/6 (cf. Table 3), which in turn performed worse than grade levels 7/8 and 10/11, whereas the means of the latter two grade levels did not differ significantly (rank order: grade level 9 < 5/6 < 7/8 = 10/11). Comparisons between the latent means of the measures of knowledge application scores showed that, once again, grade level 9 performed worst, followed by grade levels 5/6 and 10/11, neither of which differed significantly from grade level 9. Grade level 7/8 performed better than all other grade levels (rank order: grade level 9 < 5/6 = 10/11 < 7/8).

g. Similar to MicroDYN, grade level 9 had significantly lower means on CFT scores compared to grade level 7/8 ($M_{7/8} = 0$; $M_9 = -.55$, $SE = .15$, $p < .01$) and also compared to grade level 10 ($M_{10} = 0$; $M_9 = -.38$, $SE = .17$, $p < .05$), whereas the latter did not differ significantly from grade level 7/8 ($M_{7/8} = 0$; $M_{10} = -.15$, $SE = .12$, $p > .05$). The overall order of the means was comparable to the pattern for measures of knowledge acquisition (rank order: grade level 9 < 7/8 = 10; the CFT was not administered to Grades 5, 6, and 11).

In summary, findings were not as straightforward as expected because performance on all measures did not increase consistently in higher grade levels. In addition to the generally low performance in grade level 9 on all measures, measures of knowledge application scores dropped for grade level 10/11 compared to grade level 7/8, whereas measures of knowledge acquisition remained stable. Thus, Hypothesis 3 was only partially supported.

Hypothesis 4: Construct Validity

All analyses to test relations between CPS and GMA (Hypothesis 4a) used models with latent variables within structural equation modeling. However, results for CPS, GMA, GPA, and parental education (Hypotheses 4b and 4c) were based on path analyses using manifest

variables because the sample sizes of the subsamples (e.g., Hypothesis 4b: $N = 75$ in Grade 11) were not appropriate for latent analyses.

CPS and g. We assumed that g would predict CPS performance; however, a significant amount of variance was expected to remain unexplained (Hypothesis 4a). Thus, by using structural equation modeling, we regressed MicroDYN on the CFT and estimated the proportion of explained variance in the MicroDYN dimensions. Results illustrated in Table 4 showed that the CFT explained performance in measures of knowledge acquisition and knowledge application in the overall model, as well as in all separate grade level models.

Please insert Table 4 about here

Although the CFT significantly predicted performance for both dimensions, the residuals of measures of knowledge acquisition and knowledge application were still highly correlated ($r = .32$ - $.62$), indicating common aspects of CPS dimensions separable from g. The model fit well for the overall sample (CFI = .948, TLI = .971, RMSEA = .053) and showed a good to acceptable fit for the several grade level models (CFI = .932 - .992, TLI = .960 - .994, RMSEA = .032 - .062). Except for grade level 9 ($p < .01$), path coefficients of the CFT predicting the dimensions acquisition and application (left part of Table 4) differed only marginally between grade levels ($p > .05$).

Overall, participants in grade level 9 showed unexpected data patterns for Hypotheses 2, 3, and 4a: They scored the worst by far on MicroDYN and the CFT, in comparison to both other grade levels and the CFT scaling sample. Further, measures of knowledge acquisition and knowledge application were extremely highly correlated in grade level 9 (see results in Hypothesis 2). Also, MicroDYN and the CFT were related more strongly than in all other grade levels (see Hypothesis 4a and residual correlations in Table 4). The combination of poor performance on all measures and increased correlations between them indicate that covariates strongly influenced performance scores. Thus, we decided to elaborate on possible

reasons for the unexpected pattern of results in the Discussion section and to exclude grade level 9 from further analyses.

CPS, g, and GPA. Having shown that MicroDYN had a significant amount of unshared variance with the CFT, the two constructs might also differ in their predictive validity, further indicating that CPS is separable from g. Thus, we checked the incremental validity of MicroDYN beyond the CFT in predicting performance in GPA (Hypothesis 4b). We decided to use grades (e.g., Grades 7 and 8, respectively) instead of grade levels (e.g., grade level 7/8) in these analyses because school GPA is not comparable between different grades, and the same GPA in different grades (e.g., Grade 7 or Grade 8) reflects different levels of performance.

Whereas scores on MicroDYN and the CFT were based on the same test for all students, GPA depended on demands that varied across grades. We used manifest path analyses due to the small sample sizes within each grade: As shown in Table 5, the criterion GPA was predicted by only MicroDYN, only CFT, and, in a final step, by MicroDYN and the CFT simultaneously. In the last model, both predictors, the CFT and MicroDYN, were combined to determine the incremental validity of MicroDYN by comparing the explained variance of this model with the explained variance of the model containing only the CFT (indicated by ΔR^2 in Table 5).

Please insert Table 5 about here

Results displayed in Table 5 showed that although MicroDYN predicted performance in GPA, the CFT was more strongly related to GPA. Additionally, MicroDYN added a small percentage of variance when predicting GPA together with the CFT in Grades 8 and 10. Global model fit was good (RMSEA = .000 - .001, CFI = .991 - .999). Thus, Hypothesis 4b was supported even though this finding was not consistent across all grades.

CPS, g, and parental education. To investigate the impact of potential determinants of CPS, we expected that parental education would predict performance for MicroDYN and the

CFT (Hypothesis 4c). We used path analysis due to the small sample sizes within each grade and predicted performance in MicroDYN and the CFT by parental education. Results showed that mothers' education predicted performance in MicroDYN in Grade 7 ($R^2_{\text{MicroDYN}} = .03$, $p < .05$; $R^2_{\text{CFT}} = .00$, $p > .05$) and Grade 8 ($R^2_{\text{MicroDYN}} = .06$, $p < .05$; $R^2_{\text{CFT}} = .03$, $p > .05$), but not on the CFT. The opposite was true in Grade 10 ($R^2_{\text{MicroDYN}} = .00$, $p > .05$; $R^2_{\text{CFT}} = .04$, $p < .05$). Fathers' education yielded significant paths for MicroDYN and the CFT only in Grade 7 ($R^2_{\text{MicroDYN}} = .02$, $p < .05$; $R^2_{\text{CFT}} = .02$, $p < .05$), although fathers' education was significantly correlated with mothers' education ($r = .54$, $p < .01$). In summary, mothers' education predicted performance in MicroDYN and on the CFT, even though this finding was not consistent across all grades, partially supporting Hypothesis 4c.

Discussion

The aim of the present study was to enhance the understanding of complex problem solving and to evaluate its relevance in educational contexts by defining the concept and by establishing construct validity in a sample of Hungarian high school students. Generally, the results of the current study provided support for an understanding of CPS as a broad mental process measurable by means of computer-based assessment with high relevance to education. More specifically, (a) CPS was best modeled as a 2-dimensional construct with knowledge acquisition and knowledge application, (b) measurement of these two dimensions was invariant across groups composed of Hungarian high school students ranging from 11 to 17 years in age, and (c) latent mean comparisons revealed an increase in knowledge acquisition and in knowledge application in part (i.e., only from grade level 5/6 to 7/8) with increasing grade level. However, this was not true for students in Grade 9, who performed lowest on both dimensions as will be discussed later on. (d) CPS was correlated with and yet clearly distinct from a measure of g and exhibited predictive validity beyond it. Further, level of parental education was related to CPS and g yielding overall important educational implications for the understanding of complex cognitive abilities such as CPS.

(1) Dimensionality: Knowledge Acquisition and Knowledge Application

The data showed the best fit to the model that assumed the existence of two dimensions of CPS, knowledge acquisition and knowledge application. This finding supports a common assumption that knowledge acquisition is a necessary but not a sufficient condition for knowledge application. For instance, Newell and Simon (1972) stated that goal-oriented problem solving necessitates an adequate problem space, in which important knowledge about the problem is stored. However, they also acknowledged that generating and applying a solution depends on additional procedural abilities, such as forecasting, strategic planning, or carrying out planned actions (Raven, 2000). Consequently, research on CPS has generally applied a knowledge acquisition and a subsequent knowledge application phase (e.g., Kröner et al., 2005). Results in this study supported these findings within a psychometric assessment approach for different grade levels of students.

Usually ability assessment is limited to the evaluation of final solutions. That is, the final results of cognitive processes, for instance, knowledge application scores in CPS, are used in educational contexts to make selection decisions, to initiate specific training measures, or to assess an entire educational system. However, the cognitive process of deriving a representation and actually carrying out a problem solution is often disregarded, but some added value is to be expected by establishing more process-oriented measures. Clearly, CPS with its broad components is a valid candidate for such an enterprise, and future research should attend to the issue of process measures as their added value becomes available through computer-based assessment.

(2) Measurement Invariance across Grade Levels (Structural Stability)

Comparing CPS scores between grade levels requires that the assessment instrument, MicroDYN, measures exactly the same construct across groups as indicated by measurement invariance. The current study tested CPS for strong invariance of a first-order structure composed of the two dimensions knowledge acquisition and knowledge application.

According to Byrne and Stewart (2006), evidence of invariance can be based on either a traditional perspective by evaluating significant drops in overall fit or on a more practical perspective by evaluating absolute changes in fit criteria. As portrayed in Table 2, results from either perspective strongly supported the invariance of CPS across grade levels 5 through 11 in Hungarian students, speaking generally well for the MicroDYN measure and its Hungarian adoption. That is, individual differences in factor scores are due to differences in underlying ability, allowing direct comparisons of ability levels between students and between grades.

Results of tests of measurement invariance can also provide insight into the structural development and the structural stability of knowledge acquisition and knowledge application even though these are somewhat limited by the cross-sectional nature of the data. Whereas no studies have addressed the issue of structural stability in CPS until now, much is known about it in g. A large body of studies has suggested that both g on stratum III and broad cognitive abilities on stratum II within the CHC theory are shaped by the time students begin attending school (e.g., Salthouse & Davis, 2008). That is, the factorial structure of g is built early in childhood (no later than by the age of 6) and then remains constant for several decades. It is only in older age that differentiation may once again decrease, as indicated by increasing correlations among stratum II abilities and higher factor loadings on g (Deary, Whalley, & Crawford, 2004). CPS is composed of complex mental operations (Funke, 2010). Thus, differentiation into knowledge acquisition and knowledge application is unlikely to take place earlier than it takes place in g. As strict factorial invariance holds from grade levels 5/6 (youngest age: 11 years) to 10/11, this differentiation cannot take place before the age of 6 but has largely taken place by the age of 11. That is, the results of our study suggest that at the age of 11, the structural stability of CPS can be assumed.

(3) Latent Mean Comparisons across Grade Levels

After finding evidence of an invariant factor structure, the study tested latent mean differences between grade levels. Results revealed that the mean scores of grade level 7/8 were higher than those of grade level 5/6, whereas grade level 9 scored lowest on both indicators. Grade level 10/11 showed the same performance as grade level 7/8 in knowledge acquisition but showed a small and yet significant decrease in knowledge application.

Not entirely unexpected, latent scores of students in grade level 9 exhibited a substantial drop in performance on both dimensions and, additionally, on latent scores of the CFT. This drop and the consolidation of performance in grade 10/11 can be seen in the context of the transition from elementary to secondary school in Hungary, which takes place just before entering Grade 9. School transitions in general yield personal and academic challenges and are highly likely to be associated with achievement loss (e.g., Smith, 2006). In the specific case of Hungary, Molnár and Csapó (2007) showed that this performance decrease is not limited to our sample reporting a general decrease in test scores in Grade 9 for Hungarian students. These drops in academic performance tend to recover to their pretransitional level in the year following the transition (Alspaugh & Harting, 1995).

There is a mutual understanding among researchers that transition impairs achievement. However, little is known about the underlying mechanisms. Besides stress imposed by the distracting nature of changing peer relationships, new norms, and harsher grading compared to elementary school (Alspaugh & Harting, 1995), a general loss of motivation partly attributable to effects of pubertal changes (Wigfield, Byrnes, & Eccles, 2006) is assumed to further attenuate test performance (Smith, 2006). In our study, not only was mean performance level higher, but latent correlations between knowledge application, knowledge acquisition, and g were also strikingly higher in grade level 9 than in any other grade level, possibly pointing to motivational issues as the underlying cause. That is, as students were less motivated to perform well on any of the tests, the variance in performance scores was largely

generated by different levels of motivation, resulting in high correlations between constructs. This is a well-known effect in research on the development of intelligence. However, alternative explanations for the performance drop in Grade 9 are feasible as well. For instance, students in lower grades might have perceived the CPS task as some kind of game and enjoyed working on it, whereas tasks might have been simplistic and boring to students in higher grades.

Considering the significant drop precisely at the change from elementary to secondary school and the (partial) recovery in scores in grade level 10/11 at some point after the transition observed in our study, transition apparently plays a role in explaining performance patterns across grades. However, to reveal the underlying causes and to decide between competing explanations, more comprehensive and experimental studies are required. Therefore, we decided not to interpret the results from students in grade level 9 and to interpret results from grade level 10/11 with caution in all further analyses.

After excluding grade level 9, a more consistent picture of latent means could be drawn. First, scores increased significantly from grade level 5/6 to 7/8 for both CPS processes and g, showing a combined effect of school- and out-of-school experiences, even literature acknowledges that schooling plays a large role in this development (Rutter & Maughan, 2002). Substantive interpretation of these results suggests that a change in mean scores may indeed reflect true between-grade-level differences, which is in line with research that has reported that substantial cognitive development takes place at this age (Byrnes, 2001).

However, the picture is different for the change in latent means from grade level 7/8 to 10/11: whereas g and knowledge acquisition remained at least stable, there was a statistically significant albeit small drop in performance for knowledge application. This is in contradiction to Byrnes (2001), who claims that both declarative and procedural knowledge increase with age during the adolescent period without having studies including CPS available for his review. Further, the performance decrease in knowledge application from grade level

7/8 to 10/11 of the two CPS processes was accompanied by decreasing latent correlations.[3] That is, as knowledge acquisition and knowledge application exhibited different patterns of latent means across grade levels, they also became continuously less connected (shared variance dropped from 73% to 46%).

The potentially different developmental trajectories of knowledge acquisition and knowledge application and the change in correlation patterns in higher grades cannot be explained only as an effect of transition and its consequences because no drop from grade 7/8 to grade 10/11 was observed for knowledge acquisition, but rather only for knowledge application. Thus, there may be other causes that underlie this effect. This finding is in line with Spearman's (1927) law of diminishing returns, which claims that correlations between different tests decrease with increasing age, postulating a successive differentiation as time goes by. This conception has received considerable criticism from intelligence researchers, but has not been considered for CPS. One possible explanation is that the development of knowledge application and knowledge acquisition may increasingly diverge across the lifespan, similar to what Spearman (1927) proposed for g, and as our data tentatively suggest.

Another explanation for the different development trajectories of knowledge acquisition and knowledge application is that the Hungarian school system is known as a traditional system with little emphasis on procedural knowledge as captured in knowledge application (Nagy, 2008). As a consequence, knowledge application skills might have deteriorated between grade levels 7/8 and 10/11, whereas knowledge acquisition and g were at least consolidated on a stable level. Clearly, these tentative results based on cross-sectional data have to be cautiously interpreted, and other interpretations may account equally well for the different development of the two dimensions knowledge acquisition and knowledge application. Thus, replications of these results are needed as this is the first study on the development of CPS, but these findings point out interesting paths for future research.

[3] Please note that single latent correlations may differ without compromising strong measurement invariance and do not contradict the finding of invariance (Byrne & Stewart, 2006).

(4) Construct Validity: CPS, g, and External Variables

To shed further light on CPS and to relate it to other measures of cognitive performance, we investigated relations between CPS and g, GPA, and parental education. The most comprehensive and most widely acknowledged approach to understanding mental ability is found in the CHC theory, which assumes three hierarchically arranged strata of mental abilities with g located on a general stratum III (McGrew, 2009). Two questions about CPS and CHC theory need to be answered: How does CPS relate to g? And how does CPS relate to the broad cognitive abilities on stratum II?

Clearly, CPS is influenced by g (e.g., Kröner et al., 2005; Wüstenberg et al., 2012), but the path coefficients between g and CPS that ranged from .32 to .62 in this study were substantially lower than those usually reported between g and other stratum II abilities. Does this imply that CPS cannot be subsumed within stratum II? We did not explicitly measure stratum II abilities, but used the CFT to test fluid intelligence, which is assumed to be at the core of g (Carroll, 2003). In fact, fluid intelligence exhibits the highest factor loading on g, and some researchers suggest isomorphism between the two (e.g., Gustafsson, 1984). Considering that CPS is measured by dynamic and interactive tasks, whereas stratum II abilities are exclusively measured by static tasks, which do not assess the ability to actively integrate information or to use dynamically given feedback to adjust behavior (Wüstenberg et al., 2012), CPS may indeed constitute one aspect of g that is not yet included within stratum II. This may particularly hold for knowledge application, which exhibited lower correlations with g than knowledge acquisition.

Sound measures of CPS have emerged only recently and were not available in studies that have tested the CHC theory. However, new stratum II abilities, such as general knowledge or psychomotor speed, have been tentatively identified (McGrew, 2009) and have led to adaptations of the CHC theory. Further widening the view by including dynamic measures of CPS in future studies as recently proposed by Wüstenberg et al. (2012) may turn

out to increase the understanding of how mental ability is structured. Results in the current study, albeit tentative, suggest divergent validity between measures of g and CPS, even though the theoretical implications of these findings are not conclusive. On the other hand, if CPS is really important and contributes to the explanation of students' performance in educational contexts, this should be reflected by the prediction of relevant external variables.

To test this assumption, we related g and CPS to GPA and checked whether CPS incrementally predicted GPA beyond g. We further related CPS to another relevant external variable, parental education. GPA is assumed to reflect the level of academic achievement over a longer period of time and was strongly related to g in our study. This is in alignment with a large body of research and is insofar not surprising as measures of g were originally constructed to predict academic performance in school (Jensen, 1998). In addition to g, representation of complex problems indicated by knowledge acquisition added a small percentage of explained variance, whereas the paths for knowledge application were mostly not substantial. Again, this was not surprising because the representation of acquired knowledge is demanded in school more frequently than actively carrying out a pattern of solution steps (Lynch & Macbeth, 1998). Further, this pattern of results is in line with a recent study by Wüstenberg et al. (2012), who also reported the empirical significance of knowledge acquisition beyond measures of g in predicting GPA.

Parental education, which served as a predictor of both CPS and g in our study, has been shown to be the most important socioeconomic factor in influencing school performance (Myrberg & Rosen, 2008) and to be somewhat related to g. To this end, Rindermann, Flores-Mendoza, and Mansur-Alves (2010) reported a small yet significant relation of parental education and g. In our study, parental education predicted g as well as CPS, even though not consistently in all grades. One explanation for the significant relation between CPS and parental education especially in earlier grades may be that parents with higher levels of education provide more stimulating and activating learning environments, offer more

emotional warmth, and often engage in playful and educational activities with their children (Davis-Kean, 2005). These children may be confronted more often with dynamic and interactive situations, which are fundamental for acquiring and applying new knowledge.

How can these findings further inform a theoretical understanding of g, CPS, and their reciprocal relation? Clearly, g is a good predictor of academic achievement, which can be somewhat complemented by CPS as shown in this study and in Wüstenberg et al. (2012). Additional support for the relevance of CPS is found in Danner et al. (2011), who reported that CPS predicted supervisor ratings on the job beyond g. In summary, more research on the nature of CPS is needed to bolster the results found in this study, but the increase in the accuracy yielded by CPS in predicting relevant external criteria is a promising starting point.

Limitations

Obvious limitations of this study that require consideration refer primarily to sample characteristics and methodological issues: a cross-sectional design of a limited age span in only a few grade levels was used, thus prohibiting generalization of results and causal conclusions. Further, there might have been small flaws in the representativeness of our subsample, paired with potentially influential transition effects, and leading to the exclusion of Grade 9 in the analyses on construct validity. We clearly acknowledge that relations between constructs may differ depending on the methods applied (e.g., Myrberg & Rosen, 2008) and that, therefore, our results are to a certain extent tentative and not generalizable. However, a more severe problem research on CPS suffers from is due to the fact that few studies have addressed the issue of the assessment and construct validity of CPS. Thus, directly comparing our results to previous research is difficult and interpretations remain inconclusive. Clearly, research will strongly benefit from widening the view to other designs.

A second point relates to the understanding of g in this study. By employing the CFT, we tested a rather narrow aspect of g, and it is difficult to relate CPS and the CHC theory when only single measures are applied. On the other hand, fluid intelligence is the strongest

marker of g (Carroll, 2003) and one of its most frequently used tests. We suggest for further research to again widen the view by explicitly assessing different stratum II abilities. However, just as our measure of g could be challenged, this is also true for the measure of CPS: The nature of the tasks we used heavily influenced the problem solving process and narrowed it down to a certain extent, an issue faced by any latent construct. For instance, Newell and Simon (1972) suggested that problem solvers refer back to the problem space when carrying out a problem solution. This interaction between knowledge acquisition and knowledge application was not included in our study. On the other hand, the two main processes identified by problem-solving research (i.e., representation and solution) are theoretically implemented in our measure of CPS and were empirically separable. Further, careful attempts to develop CPS measures have been scarce until now, and our results suggest that using multiple complex tasks is a valid approach for capturing CPS performance.

Implications and Conclusion

The general impact of schooling on mental ability has been widely acknowledged (Rutter & Maughan, 2002). At the same time, enhancing cognitive performance in school, or in other words, improving students' minds, is a major challenge of education and an educational goal in itself (Mayer & Wittrock, 2006). In fact, large-scale assessments such as PISA are explicitly aimed at describing and comparing levels of achievement in different educational systems, but the implicitly made notion is to find ways to make education more efficient, for example, by enhancing complex cognitions such as problem solving. When it comes to these complex cognitions, it is often assumed that this challenge is met implicitly in school. To describe this phenomenon, Mayer and Wittrock (2006) introduced the term *hidden curriculum*, stating that "educators expect students to be able to solve problems […] but rarely provide problem-solving instruction" (p. 296). The assumption of a hidden curriculum may partly be unjustified as the results of our study suggest: CPS and its components were not as strongly fostered as one might have hoped.

In the search for methods that promote CPS, Mayer and Wittrock (2006) listed seven instructional methods with a more or less proven impact on problem solving. However, one general disadvantage of approaches aimed at enhancing problem-solving skills is that evidence for transfer to other types of problems is rather scarce (Mansfield, Busse, & Krepelka, 1978). To this end, Mayer and Wittrock (2006) concluded that teaching domain-specific skills is more promising than trying to foster domain-general CPS abilities.

At this point, we are less pessimistic and differ in our view from Mayer and Wittrock (2006). Similar to our position, Novick, Hurley, and Francis (1999) underline the importance of general processes in problem solving by stating that abstract representation schemas (e.g., causal models or concept maps) are more useful than specifically relevant example problems for understanding the structure of novel problems because these general representations are not contaminated by specific content (Holyoak, 1985). Also Chen and Klahr (1999) showed that training students in how to conduct experiments that allow for causal inferences led to an increase in the knowledge acquired, even though it was gathered in a specific context (i.e., science education). This knowledge was successfully transferred to different tasks. Specifically, students in the experimental group performed better on tasks that were comparable to the original one but also in generalizing the knowledge gained across various tasks (Chen & Klahr, 1999).

In line with Chen and Klahr (1999), the results of our study also support the concept of generally important and transferable CPS processes. Changes in students' CPS performance may very well be reflected by corresponding increases in MicroDYN scores, independent of whether they are induced by specific training methods such as guided discovery or by school in general. Therefore, we suggest thoroughly investigating the educational implications of using MicroDYN as a training tool for domain-unspecific knowledge acquisition and application skills. It is under this assumption that CPS is employed in PISA 2012 as a domain-general and complementary measure to domain-specific concepts (OECD, 2010).

However, even though today's students need to be prepared to meet different challenges than 30 years ago, and the concept of life-long learning, thus extending the educational span to a lifetime, has become increasingly popular (Smith & Reio, 2006), one should not count one's chickens before they hatch. Said otherwise, it may be premature to consider specific training issues. Further, a deeper understanding of CPS and its relation to g seems to be needed in light of the scarce empirical evidence. With the present study, we want to empirically and conceptually contribute to this new debate, and we conclude by emphasizing the great potential that CPS has as an educationally relevant construct. Just as Alfred Binet and Théodore Simon (1905) saw the relevance of general mental ability for academic achievement and laid the foundation of modern intelligence research, Gestalt psychologists such as Karl Duncker (1945) were well aware of the implications and importance of problem solving in education. However, it is only in light of current developments that the issue of how to make students good problem solvers is finally receiving the attention it deserves within psychology.

References

Alspaugh, J. W. & Harting, R. D. (1995). Transition effects of school grade-level organization on student achievement. *Journal of Research and Development in Education, 28*(3), 145-149.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing (3rd edition).* Washington, DC: AERA.

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale – Fourth Edition: What does the WAIS-IV measure? *Psychological Assessment, 22*(1), 121-130.

Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anomaux [New methods for assessing the intellectual level of anormal individuals]. *L'Année Psychologique, 11*, 191-244.

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

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.

Byrne, B. M. & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*(2), 287-321.

Byrnes, J. P. (2001). *Minds, brains, and education.* New York, NJ: Guilford 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.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5-21). Amsterdam, NL: Pergamon.

Chen, H. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.

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.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255.

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.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement. The indirect role of parental expectations and the home environment. *Journal of Family Psychology, 19*(2), 294-304.

Deary, I. J., Whalley, L. J., & Crawford, J. R. (2004). An "instantaneous" estimate of a lifetime's cognitive change. *Intelligence, 32,* 113-119.

Diaz, L. & Heining-Boynton, A. L. (1995). Multiple intelligences, multiculturalism, and the teaching of culture. *International Journal of Educational Research, 23,* 607-617.

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

Dörner, D. (1990). The logic of failure. In D. E. Broadbent, J. T. Reason, & A. D. Baddeley (Eds.), *Human factors n hazardous situations* (pp. 15-36). New York, NY: Oxford University Press.

Dörner, D. & Kreuzig, W. (1983). Problemlösefähigkeit und Intelligenz [Problem solving ability and intelligence]. *Psychologische Rundschau, 34,* 185-192.

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

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

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

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

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

Gardner, P. H. & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9*, 55-79.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ.* New York, NY: Bantam Books.

Greeno, J. G., & Simon, H. A. (1988). Problem solving and reasoning. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (pp. 239-270). Hillsdale, NJ: Erlbaum.

Greiff, S. (2012). *Individualdiagnostik der Problemlösefähigkeit* [Diagnostics of problem solving ability on an individual level]. Münster: Waxmann.

Greiff, S. (in press). From Interactive to Collaborative Problem Solving: Current issues in the Programme for International Student Assessment. *Review of Psychology.*

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement, 36*(3), 189-213.

Gustafsson, J. E. (1984). An unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179-203.

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.

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.

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

Kihlstrom, J. F. & Cantor, N. (2011). Social intelligence. In R. J. Sternberg & S. C. Barry, *The Cambridge handbook of intelligence* (pp. 564-581). New York, NY: Cambridge University press.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*(4), 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 – Lessons learned from large-scale surveys and implications for testing* (pp. 151-156). Luxembourg: Office for Official Publications of the European Communities.

Lievens, F. & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J. L. Farr & N. T. Tippins (Eds), *Handbook of employee selection* (pp. 339-359). New York, NY: Routledge.

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). Mahwah, NJ: Lawrence Erlbaum.

Mansfield, R. S., Busse, T. V., & Krepelka, E. J. (1978). The effectiveness of creativity training. *Review of Educational Research, 48*, 517-536.

Markman, A. B. (1999). *Knowledge representation.* Mahwah, NJ: Erlbaum.

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.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10.

Molnár, G., & Csapó, B. (2007). *Constructing complex problem solving competency scales by IRT models using data of different age groups.* Abstract submitted at the 12th Biennial Conference of EARLI 2007, Budapest, Hungary.

Muthén, L.K. & Muthén, B.O. (2010). *Mplus User's Guide* (6th edition). Los Angeles, CA: Muthén & Muthén.

Myrberg, E. & Rosen, E. (2008). A path model with mediating factors of partens' education on students' reading achievement in seven countries. *Educational Research and Evaluation, 14* (6), 507-520.

Nagy, J. (2008). Renewing elementary education. In K. Fazekas, J. Köllö, & J. Varga (Eds.), *Green book for renewal of public education in Hungary* (pp. 61-80). Budapest: Ecostat.

Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: knowns and unknowns. *American Psychologist, 51*, 77-101.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

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

Novick, L. R., Hurley, S. M., & Francis, M. (1999). Evidence for abstract, schematic knowledge of three spatial diagram representation. *Memory and Cognition, 27*, 288-308.

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

OECD (2009). *PISA 2009 results: What students know and can do: Student performance in Reading, Mathematics, and Science (Volume I).* Paris: OECD.

OECD (2010). *PISA 2012 problem solving framework (draft for field trial).* Paris: OECD.

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

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

Raven, J. C. (1962). *Advanced progressive matrices, Set II*. London: Lewis.

Ree, H. M. & Carretta, T. R. (2002). g2K. *Human Performance, 15,* 3-23.

Reeve, C. L. & Hakel, M. D. (2002). Asking the right questions about g. *Human Performance, 15,* 47-74.

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

Rindermann, H., Flores-Mendoza, C., & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents' socioeconomic status and education. *Learning and Individual Differences, 20*(5), 544-548.

Rost, D. H. (2009). *Intelligenz* [Intelligence]. Weinheim: PVU Beltz.

Rutter, M. & Maughan, B. (2002). School effectiveness findings 1979-2002. *Journal of School Psychology, 40*(6), 451-475.

Salthouse, T. A. & Davis, H. P. (2008). Organization of cognitive abilities and neuropsychological variables across the lifespan. *Developmental Review, 26*, 31-54.

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.

Smith, S. S. (2006). Examining the long-term impact of achievement loss during the transition to high school. *The Journal of Secondary Gifted Education, 17* (4), 211-221.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab. Acceptance and psychometric characteristics of a computer-based microworld to assess Complex Problem Solving. *Psychological Test and Assessment Modeling, 54*(1), 54-72.

Spearman, C. (1927). *The abilities of man. Their nature and measurement.* New York, NY: Macmillan.

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* (p. 295-321). Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J. (2000). The holey grail of general intelligence. *Science, 289,* 399-401.

Sternberg, R. J. (2009). Toward a triachic theory of human intelligence. In R. J. Sternberg, J. C. Kaufman, & E. L. Grigorenko (Eds.), *The essential Sternberg: Essays on intelligence, psychology, and education* (pp. 38-70). New York, NY: Springer.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement Issues and Practice, 14*(3), 29-36.

Taub, G. E. & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly, 19*(1), 72-87.

Ullman, J. B. (2007). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 676-780). Boston, MA: Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale. 4th edition.* San Antonio, TX: Pearson Asssessment.

Weiß, R. H. (2006). Grundintelligenztest Skala 2 - Revision - (CFT 20-R) [Culture Fair Intelligence Test 20-R – Scale 2]. Göttingen: Hogrefe.

Wigfield, A., Byrnes, J. P., & Eccles, J. S. (2006). Development during early and middle adolescence. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 87-114). Mahwah, NJ: Lawrence Erlbaum.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex Problem Solving. More than reasoning? *Intelligence*, *40,* 1-14.

Table 1

Goodness of Fit Indices for Testing Dimensionality of MicroDYN, Overall and by Class Level

Model	χ^2	*df*	*p*	CFI	TLI	RMSEA	n
2-dimensional including all class levels	164.068	53	.001	.967	.978	.050	855
1-dimensional including all class levels	329.352	52	.001	.912	.944	.079	855
2-dimensional only class level 5/6	65.912	35	.001	.966	.966	.064	216
2-dimensional only class level 7/8	77.539	13	.001	.969	.969	.056	300
2-dimensional only class level 9	13.908	29	.380	.996	.996	.029	83
2-dimensional only class level 10/11	51.338	40	.001	.991	.991	.033	256

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.

Table 2

Goodness of Fit Indices for Measurement Invariance of MicroDYN

Model	χ^2	*df*	*Compare with*	$\Delta\chi^{2\,(1)}$	$\Delta df^{(1)}$	p	CFI	TLI	RMSEA
(1) Configural Invariance	161.045	104					.975	.975	.051
(2) Strong Factorial Invariance	170.101	115	(1)	22.294	23	>.10	.976	.982	.047
(3) Strict Factorial Invariance	165.826	116	(1)	53.159	43	>.10	.978	.983	.045

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

[1)] χ^2 and Δdf were estimated by Difference Test-procedure in MPlus (see Muthén & Muthén, 2004). χ^2- differences between models cannot be compared by substracting χ^2 and *df*, if WLSMV-estimators are used.

Table 3

Latent Mean Comparisons of Knowledge Acquisition and Knowledge Application (MicroDYN) Between Different Class Levels

Dimension	Model	*Compare with*	*M (SE)*	p	*Compare with*	*M (SE)*	p	*Compare with*	*M (SE)*	p
Acquisition	(1) classlevel 5+6		.00							
	(2) classlevel 7+8	(1)	.18 (.11)	<.05						
	(3) classlevel 9	(1)	-.37 (.17)	<.05	(2)	-.54 (.15)	<.001			
	(4) classlevel 10+11	(1)	.30 (.15)	<.05	(2)	.04 (.13)	>.05	(3)	.88 (.36)	<.01
Application	(1) classlevel 5+6		.00							
	(2) classlevel 7+8	(1)	.50 (.24)	<.05						
	(3) classlevel 9	(1)	-.52 (.25)	<.05	(2)	-.72 (.29)	<.001			
	(4) classlevel 10+11	(1)	.04 (.13)	>.05	(2)	-.24 (.11)	<.05	(3)	.88 (.36)	<.01

Note. M = Latent Mean; SE = Standard Error. Statistical significance of the differences between all groups was determined by z-statistics.

Table 4

Prediction of Performance in Knowledge Aquisition and Knowledge Application (MicroDYN) by g, Overall and by Class Level

Model	*path coefficient acquisition*	*path coefficient application*	R^2 *acquisition*	R^2 *application*	*residual correlation aquisition / application*	*N*
Overall	.47 (.04)***	.40 (.05)***	.22 (.04)***	.16 (.04)***	.63 (.05)***	486
Class level 7/8	.48 (.05)***	.39 (.07)***	.23 (.05)***	.15 (.05)**	.60 (.06)***	284
Class level 9	.62 (.12)***	.62 (.12)***	.38 (.14)**	.38 (.15)**	.30 (.10)***	79
Class level 10	.34 (.10)***	.32 (.11)***	.11 (.07)*	.11 (.07)*	.62 (.08)***	123

*p < .05. **p < .01. ***p < .001.

Table 5

Prediction of GPA by MicroDYN and CFT

Class	R^2 in GPA				
	MicroDYN	CFT	MicroDYN &CFT	Δ R^2	N
7	.03*	.19***	.19***	.00	104
8	.08*	.09***	.13***	.04*	93
10	.07*	.15***	.18***	.03*	90
11	.07*				75

Note. R^2 = explained variance. ΔR^2 = "*" indicates significant path coefficients of CPS contributing to R^2.

*p < .05. **p < .01. ***p < .001.

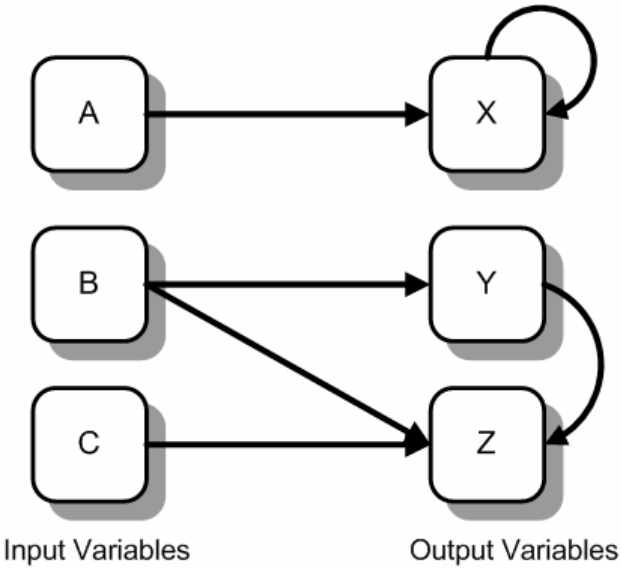

Figure 1. Structure of a typical MicroDYN task displaying three input (A, B, C) and three output (X, Y, Z) variables.

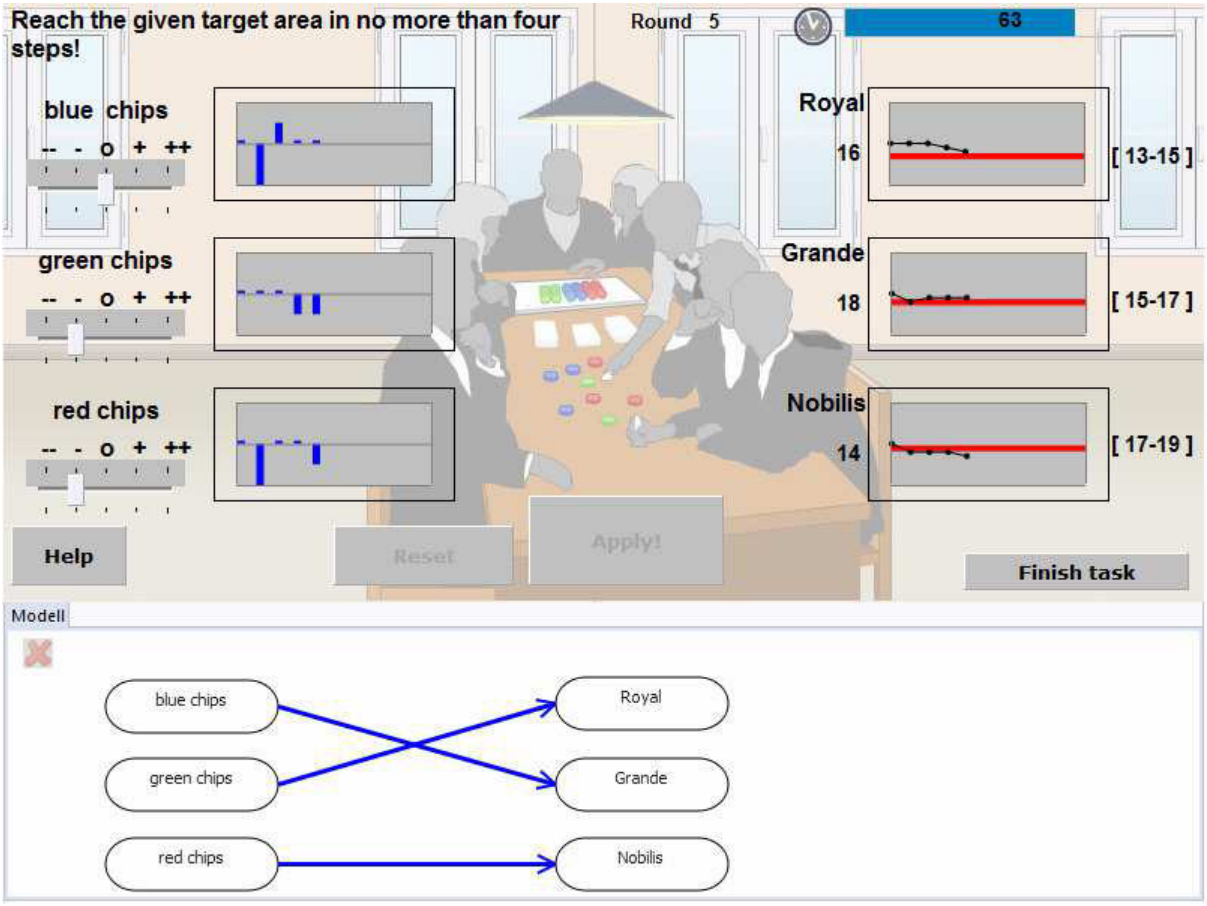

Figure 2. Screenshot of the MicroDYN task "*Game Night.*" The controllers of the input variables range from "- -" (value = -2) to "++" (value = +2). The current values and the target values of the output variables are displayed numerically (e.g., current value for Royal: 16; target values: 13-15) and graphically (current value: dots; target value: red line). The correct model is shown at the bottom of the figure.

Appendix

The seven items in this study were varied mainly on two system attributes proven to be most influential on item difficulty (see Greiff, 2012): the number of effects between variables and the quality of effects (i.e., effects of input and output variables).

	Linear Structural Equations	***System size***	***Effects***
Item 1	$X_{t+1} = 1*X_t + 0*A_t + 2*B_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t$	2x2-System	only direct
Item 2	$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2x3-System	only direct
Item 3	$X_{t+1} = 1*X_t + 0*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only direct
Item 4	$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 2*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only direct
Item 5	$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only direct
Item 6	$X_{t+1} = 1.33*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2x3-System	direct and indirect
Item 7	$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_{t+1} = 1.33*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-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, whereas X_{t+1}, Y_{t+1}, Z_{t+1} denote the values of the output variables in the subsequent trial.

4

Assessment with Microworlds:

Factor structure, Invariance, and Latent Mean

Comparison of the MicroDYN test.

This article is currently under review:

Greiff, S., & Wüstenberg, S. (submitted). Assessment with microworlds: factor structure, invariance, and latent mean comparison of the MicroDYN test. *European Journal of Psychological Assessment.*

Abstract

Computer-simulated microworlds have witnessed significant international interest over the last decades as assessment vehicles for complex mental skills. This interest strongly contrasts to what is currently known about measurement characteristics of microworlds. In this study factorial structure, measurement invariance, and latent means of the MicroDYN measure, a computer-based assessment instrument containing an entire set of dynamic microworlds, were examined in 4 German subsamples of junior high school students in 8th to 10th grade (N = 309), senior high school students in 11th to 13th grade (N = 484), university students (N = 222), and blue-collar workers (N = 181). The findings support a 2-dimensional structure of the MicroDYN measure with the dimensions knowledge acquisition and knowledge application, suggest satisfactory measurement invariance across all samples, and yield meaningful comparisons between latent means with university students performing best. It is suggested to further explore measurement characteristics of computer-simulated microworlds to fully exploit their potential as means of modern assessment instruments. Implications and limitations are discussed.

Assessment with Microworlds: Factor Structure, Invariance, and Latent Mean Comparison of the MicroDYN Test

The advent of computers in assessment inaugurated a new era of testing across the world: computer-based assessment (CBA) provided the means to optimize standardization of test administration, to increase test efficiency through adaptive testing, to score testees' answers automatically, to provide immediate feedback, and to gain insights into processes by using information stored in log files. Besides these advantages (cf. Scheuermann & Björnsson, 2009; Van der Linden & Glas, 2000) and associated challenges (e.g., test mode effects; Clariana & Wallance, 2002; Johnson & Green, 2006), exploiting means of technology through CBA allows for a range of new types of skills to be measured, which were not measureable by traditional means of paper-and-pencil testing (Kyllonen, 2009).

This explains the motivation behind the recent shift towards CBA in large-scale assessments such as the Programme for International Student Assessment (PISA): Already in the PISA 2006 survey, a computer-assisted assessment of electronic reading was included (OECD, 2006) and the upcoming PISA survey 2012 will use dynamic microworlds to test cognitive skills associated with interactive problem solving (OECD, 2010). In fact, computer-simulated systems (i.e., microworlds) have frequently been used in the assessment of complex mental processes testing participants' abilities to actively explore unknown systems, to use efficient meta-cognitive strategies, to acquire knowledge in new and intransparent situations, to actively apply this knowledge, and to readily adapt and respond to actions initiated by the system (Funke, 2001; Raven, 2000). Even though the degree of implemented real-world analogy in these microworlds varies considerably, they all require active action regulation such as incorporating feedback given by the system or goal setting and achievement (Kröner, Plass, & Leutner, 2005).

In the early stages of assessing performance in microworlds, a high resemblance with real-world phenomena was considered crucial and simulations required participants to act as mayor of a town (Dörner & Wearing, 1995), to manage medical systems (Gardner & Berry, 1995), or to coordinate fire fighters (Rigas, Carling, & Brehmer, 2002). However, persistent measurement issues (e.g., insufficient standardization, low reliability, obscured scoring; for an overview consult Greiff, Wüstenberg, & Funke, 2012 and Kröner et al., 2005) initiated a move towards more standardized, fictitious, and thereby less realistic computer simulations. In their microworld *Multiflux*, Kröner et al. (2005) confront participants with an artificial mechanical machine, whereas Kluge's (2008) *ColorSIM* allows testees to create different colors by mixing arbitrarily named substances. An entire set of computer-simulated microworlds with reduced time-on-task is presented within the *Genetics Lab* assessment (Sonnleitner et al., 2012) as well as in the *MicroDYN* test (Greiff et al., 2012), the latter constituting the measures used for the PISA 2012 survey (OECD, 2010).

Individual performance is usually decomposed into *knowledge acquisition* (i.e., representation of a microworld's structure; Mayer & Wittrock, 2006; Novick & Bassok, 2005) and *knowledge application* (i.e., finding a solution to reach a desired goal state; Funke, 2001; Klahr & Dunbar, 1988). This decomposition is shown to hold empirically even though both dimensions are substantially correlated (Bühner, Kröner, & Ziegler, 2008; Greiff et al., 2012; Kröner et al., 2005). Further, complex mental processes assessed in microworlds are correlated with and yet distinct from traditional measures of intelligence (Gonzalez, Thomas, & Vanyukov, 2005; Wenke, Frensch, & Funke 2005; Wüstenberg, Greiff, & Funke, 2012) and are reported to incrementally predict relevant outcomes such as school grades (Greiff & Fischer, in press; Wüstenberg et al., 2012) or supervisory ratings (Danner et al., 2011) beyond measures of intelligence. It has therefore been argued that computer-simulated microworlds are attractive candidates for the measurement of complex mental skills (Funke, 2001; Greiff et al., 2012; Kröner et al., 2005).

However, to date little is known about measurement characteristics of microworlds such as MicroDYN, in particular in different target populations. That is, notwithstanding research mentioned above, important issues such as measurement invariance are yet to be scrutinized and previous findings almost entirely rely on homogeneous samples of university students. The present study examines factor structure as well as measurement invariance, and compares latent means for the computer-simulated MicroDYN test (Greiff et al., 2012) in four samples composed of secondary school students (junior high school in 8^{th} to 10^{th} grade and senior high school in 11^{th} to 13^{th} grade), university students, and blue-collar workers.

MicroDYN and factor structure

Replicating previous findings, we expect the two skills, knowledge acquisition and knowledge application, to be substantially correlated and yet distinct from each other in the overall sample and in all four subsamples *(Hypothesis 1).*

MicroDYN and measurement invariance

A valid comparison of test scores between different populations is predicated on equivalent meaning of these scores across subgroups (French & Finch, 2006). That is, measures must exhibit a certain degree of factorial invariance, which is tested within multigroup confirmatory factor analysis (Byrne & Stewart, 2006). We expect the MicroDYN simulations to be measurement invariant across the four subsamples *(Hypothesis 2).*

MicroDYN and latent mean comparisons

Only to the extent measurement invariance holds, between-group differences can be interpreted as true and not psychometric differences in underlying skills (Cheung & Rensvold, 2002). For those parts of the measurement model, for which invariance can be assumed, latent factor means between groups are compared. No published results on latent mean comparisons in microworlds are available, but as educational, academic, and job attainment is substantially associated with cognitive performance in general (cf. Rohde & Thompson, 2007; Schmidt &

Hunter, 1998), this may also hold for complex mental processes relevant when engaging with microworlds. We therefore expect the samples of senior high school students and university students to perform better than students in junior high school and blue-collar workers. No further assumptions on performance differences are posed and no differential effects of performance patterns for knowledge acquisition and knowledge application are a priori expected *(Hypothesis 3).*

Methods

Participants

The overall sample contained four subgroups[1] of mostly German nationality including junior high school students (*N*=309; 147 female, 162 male; mean age: *M*=14.56, *SD*=1.05; grades 8 to 10), senior high school students (*N*=484; 281 female, 198 male, 5 missing sex; mean age: *M*=17.68, *SD*=1.21 grades 11 to 13), university students (*N*=222; 154 female, 66 male, 2 missing sex; mean age: *M*=22.80, *SD*=4.0) and blue-collar workers (*N*=181; 12 female, 169 male; mean age: *M*=42.85, *SD*=10.57). School education varied considerably across subgroups, including students or graduates of three different school tracks within German school system (general, intermediate, and high track). In Germany, only the high track continues after grade 10 and offers direct access to tertiary education. Within our sample, blue-collar workers had the lowest level of education (43% general track; 36% intermediate track; 21% high track), followed by junior high school students (17% general track; 51% intermediate track, 32% high track), senior high school students (100% high track) and university students (100% high track and within tertiary education). Test administration was employed within the same computer-based environment for all subgroups. Testing took place at local computer rooms of three different schools (high school students), at university

[1] Data of single subgroups are already used in other publications, mostly to determine construct and predictive validity of MicroDYN. Analyses on measurement invariance and on latent mean comparisons combining subgroups as presented in this study are fully original.

(university students), and directly at the work place (blue-collar workers). Blue-collar workers were employed at a big car manufacturer located in Southern Germany.

Measures

The fully computer-based MicroDYN test is embedded into the formal framework of linear structural equation (LSE) systems introduced by Funke (2001), which are frequently used in assessment contexts (Funke & Frensch, 2007). MicroDYN employs an entire set of independent microworlds with time-on-task being approximately 5 minutes for each microworld. The short time-on-task, which is referred to as minimal complexity in the literature (cf. Greiff et al., 2012), yields several measurement advantages and has proven useful for the computer-based assessment of complex mental skills. A comprehensive description of the MicroDYN approach can be found in Greiff et al. (2012).

A typical MicroDYN microworld contains a limited number of input variables, which are related to a number of output variables. Causal links can 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 participants' interventions to the microworld (Funke & Frensch, 2007). The connections between inputs and outputs are not visible to participants. With regard to semantic embedment, each microworld has a different cover story (e.g., creating chemical substances, growing pumpkins, or coaching a handball team). To avoid uncontrolled influences of prior knowledge, inputs and outputs are either labelled without deep semantic meaning (e.g., training A) or fictitiously (e.g., Natromic as name of a fertilizer). A screenshot of an exemplary task is displayed in Figure 1.

Please insert Figure 1 about here

The procedure in each MicroDYN microworld is identical to the one generally applied in LSE systems, in which participants perform two complex cognitive tasks, knowledge

acquisition and knowledge application. In phase 1, knowledge acquisition, participants are asked to explore the microworld by changing values of the inputs and to represent their gathered knowledge in a causal diagram (Funke, 2001) within a maximum of 3.5 minutes. In phase 2, knowledge application, respondents have to achieve given target values by adequately manipulating the microworld with no more of four active manipulation rounds and within a maximum of 1.5 minutes. Thus, indicators for knowledge acquisition and knowledge application as measures of performance are derived in phase 1 and 2, respectively.

A set of MicroDYN microworlds starts with a detailed instruction including trial tasks, in which participants learn which tasks they are expected to carry out and how to operate the software interface. It continues with several independent microworlds with different underlying structures, each of which is administered in exactly the same way. In the specific MicroDYN test used in this study, six microworlds were employed lasting overall approximately 40 minutes, instruction included. The specific linear equations of all microworlds are found in the Appendix. Each of the six MicroDYN microworlds yielded indicators on knowledge acquisition and knowledge application totaling in 12 indicators. With regard to knowledge acquisition, full credit was given if participants' models contained no wrong or additional, but all actual causal links, otherwise zero credit was assigned. A full score in knowledge application was given if goal values were reached, whereas no credit was assigned if target values were not fully reached (for details on scoring cf. Greiff et al., 2012; Kröner et al., 2005).

Statistical analyses

We used confirmatory factor analysis (CFA) to empirically evaluate decomposition of performance into knowledge acquisition and knowledge application (Hypothesis 1) and multiple group confirmatory factor analyses (MGCFA) within structural equation modeling (SEM; Bollen, 1989) to test for factorial invariance (Hypothesis 2) as well as to compare

latent means of subgroups (Hypothesis 3). Weighted Least Squares Mean and Variance adjusted (WLSMV; Muthén & Muthén, 2010) estimator for categorical outcomes was used for all analyses, which were conducted in the software package Mplus 5.0 (Muthén & Muthén, 2007).

Results

Hypothesis 1: MicroDYN and factor structure

The 2-dimensional model comprising knowledge acquisition and knowledge application showed a good fit in the overall sample and in each subgroup (Table 1).

Please insert Table 1 about here

According to Hypothesis 1, a 2-dimensional model fitted better than a 1-dimensional model indicated by a χ^2-difference test[2] (overall sample: χ^2= 86.206, *df*=1, *p*<.001; subgroups: χ^2= 15.695-28.945, *df*=1, *p*<.001), although both dimensions were substantially correlated (overall sample: r= .81; subgroups: r=.76-.84). Thus, the 2-dimensional model was used as baseline model for all subsequent analyses. Descriptive analyses of this model showed that mean task difficulty for the overall sample was moderate for both knowledge acquisition (*M*=.50; *SD*=.29) and knowledge application (*M*=.47; *SD*=.11), but varied considerably in subgroups (knowledge acquisition: *M*=.34-.65; *SD*=.26-.31; knowledge application: *M*=.35-.58; *SD*=09-.13). Cronbachs' α was acceptable (overall sample: knowledge acquisition: α=.74; knowledge application: α=.73; subgroups: knowledge acquisition: α=.66-.71; knowledge application: α=.69-.76). In summary, Hypothesis 1 was supported.

[2] Computing the differences of χ^2-values and *df*s by subtracting values of the 1-dimensional and 2-dimensional model in Table 1 to compare models is not appropriate when WLSMV-estimator is used. Thus, we chose a procedure integrated in the Software Mplus to compute χ^2–difference tests (Muthén & Muthén, 2010, p. 435).

Hypothesis 2: MicroDYN and measurement invariance

To analyze measurement invariance, we followed a specific procedure for categorical data described by Muthén and Muthén (2010, p.434). First, we tested configural invariance by estimating parameters of the baseline model in a multigroup model. That is, thresholds and factor loadings were not constrained across groups, factor means were fixed at 0 in all groups and residual variances were fixed at 1 in all groups. The configural invariance model showed a good fit (χ^2=228.771, *df*=110, *p*<.001; CFI=.968; RMSEA=.060), indicating that the same number of factors can be assumed within all subgroups (Byrne & Stewart, 2006).

Subsequently, the configural invariance model was directly compared to a model of strong factorial invariance, in which both factor loadings and thresholds were constrained to be equal across groups, residual variances were fixed at 1 in one group and freed in the others, and factor means were fixed at 0 in one group and freed in the others (Muthén & Muthén, 2010, p.434). This model of strong factorial invariance also fitted well (χ^2=269.800, *df*=121, *p*<.001; CFI=.960; RMSEA=.064), although slightly worse than the model of configural invariance. The drop in global model fit between both models (ΔCFI=.008; ΔRMSEA=.004) was below the criterion of ΔCFI=.01 proposed by Cheung and Rensvold (2002), indicating that measurement invariance holds from a "practical perspective" (Byrne & Stewart, 2006, p. 290). However, from a stricter traditional perspective, only a non-significant χ^2-difference test between the two models supports strong factorial invariance, which was not the case in this study (χ^2=57.018, *df*=18, *p*<.001). To further investigate potential reasons for invariance on item level, we analyzed which parameters contributed to the decrease in fit in the strong factorial invariance model (Byrne & Stewart, 2006).

Please insert Table 2 about here

We, therefore, conducted Lagrange-Multiplier-Tests with Bonferroni-correction within CFA, which correspond to differential item functioning analyses in Item Response Theory

(Byrne & Stewart, 2006). Results revealed that constraining knowledge acquisition parameters of the first regular task after instruction significantly decreased model fit, indicating non-invariance. This is illustrated in Table 2, in which factor loadings and thresholds of a configural invariance model without equality constraints (Model 1) are compared with a model including equality constraints (i.e., common factor loadings and thresholds; Model 2). Parameters of the first regular task differed markedly for knowledge acquisition, showing that blue-collar workers performed worse than all other groups beyond overall mean differences. Subsequent analyses with a partial invariance model, in which parameters of the first regular task were not constrained, led to a significant improvement of the strong factorial invariance model (χ^2=236.959, *df*=121, *p*<.001; CFI=.968; RMSEA=.058) differing non-significantly (χ^2-difference test=18.467, *df*=13, *p*>.10) from the model of configural invariance (χ^2=228.771, *df*=110, *p*<.001; CFI=.968; RMSEA=.060). In summary, measurement invariance was confirmed from a practical, but not from a strict traditional perspective, because the first task behaved differently across groups. Thus, Hypothesis 2 was only partly supported.

Hypothesis 3: MicroDYN and latent mean comparisons

To compare latent means across groups, we used the measurement invariance model (including the first task), imposed equality constraints on thresholds and factor loadings, and set the latent means of a reference group to zero (Muthén & Muthén, 2010). Thus, the estimated means for all other groups represent mean differences in relation to this group. Statistical significance of differences between all groups was determined by z-statistics. In the first comparison, junior high school students served as reference group (left part of Table 3), whereas blue-collar workers served as reference group in a second (middle part of Table 3) and senior high school students in a third comparison (right part of Table 3).

Please insert Table 3 about here

We expected senior high school students and university students to outperform junior high school students and blue-collar workers. Results confirmed these assumptions for both dimensions: senior high school students performed better than junior high school students and blue-collar workers, whereas university students performed best. Junior high school students and blue-collar workers did not differ significantly in either dimension (overall rank order: university > senior high school > junior high school = blue-collar workers for knowledge acquisition and knowledge application). However, effect sizes of latent mean differences were consistently larger in knowledge acquisition than in knowledge application. To evaluate stability of these results, we also estimated latent mean differences of a partial measurement invariance model, in which thresholds and factor loadings of the first task were not constrained. Absolute values of latent mean differences and effect sizes of this model were comparable to results of the measurement invariance model including constrained parameters of the first task. Thus, although MicroDYN was only measurement invariant from a practical perspective, latent mean differences based on a model assuming full measurement invariance were sufficiently robust. In summary, performance in both dimensions increased consistently with higher educational level and job attainment, supporting Hypothesis 3.

Discussion

This study was set out to compliment existing research on computer-based assessment in microworlds and on measurement issues associated with them. More specifically, the 2-dimensional structure composed of knowledge acquisition and knowledge application was shown to hold in heterogeneous subgroups, thereby replicating and extending previous findings. Measurement across these groups was sufficiently invariant, and latent means between groups ranked according to groups' mean educational level and in line with our expectations.

Whereas the six microworlds were measurement invariant from a practical perspective, the first task tended to be too difficult in the low performing blue-collar worker group from a traditional perspective (Byrne & Stewart, 2006). This is in line with Veenman and Elshout (1994) reporting performance impairments with decreasing instructional support only in participants of low cognitive level and suggests that an additional microworld during instruction may enhance measurement accuracy in low performance samples. Groups in academic tracks (i.e., senior high school and university students) performed substantially better in both knowledge acquisition and knowledge application than participants in non-academic tracks aligning our findings with between-group differences in cognitive skills in general. However, effect sizes were notably lower in knowledge application. Longitudinal studies need to address whether this gap is reduced over time as differing levels of departure are compensated by experiences gathered on the job (indicated by the blue-collar workers, the group with the lowest educational level and the highest age, but equal performance compared to junior high school students) or is less marked than performance differences in knowledge acquisition from the outset suggesting a different set of underlying cognitive processes for the two skills (Funke, 2001). Additionally, university students performed significantly better than senior high school students. This may be either due to a further selection at university entrance (our sample was composed of university students mainly in highly selective undergraduate and graduate programs) or due to enduring development of complex mental skills beyond high school in line with research on other cognitive skills showing that cognitive performance peaks during late adolescence and early adulthood and may be further enhanced by tertiary education (e.g., Asato, Sweeney, & Luna, 2006).

The findings add coherently to existing evidence of assessment with microworlds, but some limitations in this study need consideration. First of all, albeit the diversity of subgroups, these do not representatively map the structure of the underlying population of high school students, university students, and blue-collar workers limiting generalizability of these

findings. Further, participants were mostly of German nationality and data did not yield any information on cultural differences. However, cultural differences as well as ethnic affiliation are a major concern in large-scale assessments such as the PISA survey and need close consideration if microworlds are to be broadly used as vehicles of psychological assessment. A second concern besides cross-cultural issues is the validity of an educational assessment beyond the specific age at which testing is conducted and beyond specific educational levels. To this end, results reported here suggest structural stability of complex mental skills assessed in microworlds across groups of different age and educational attainment underlining the potential lying in modern and computer-based assessment with microworlds – a fact that has often been neglected. That is, psychometric characteristics of microworlds are not as black as they have often been painted and thoroughly penetrating their measurement characteristics and the underlying construct will render computer-simulated microworlds even more attractive in a number of assessment settings.

References

Asato, M. R., Sweeney, J. A., & Luna, B. (2006). Cognitive processes in the development of TOL performance. *Neuropsychologia, 44,* 2259-2269.

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

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

Byrne, B. M. & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Clariana, R. & Wallance, P. (2002). Paper-based versus computer-based assessment: key factors associated with test mode effect. *British Journal of Educational Technology, 33,* 593-602.

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.

Dörner, D. & Wearing, A. J. (1995). Complex problem solving: Toward a (computer-simulated) theory. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 65-99). Hillsdale, NJ: Erlbaum.

French, B. & Finch, W. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13,* 378-402.

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

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

Gardner, P. H. & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9*, 55-79.

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

Greiff, S. & Fischer, A. (in press). 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*.

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

Johnson, M. & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4*, 1-34.

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

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*, 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 Communitites.

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.

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

Muthén, L. K. & Muthén, B. O. (2010). *Mplus User's Guide* (6th edition). Los Angeles, CA: Muthén & Muthén.

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

OECD (2006). *The Programme for International Student Assessment 2006.* Accessed at: http://www.oecd.org/dataoecd/15/13/39725224.p*df*.

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

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

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

Rohde, T. E. & Thompson, L. A. (2007). Predicitng academic achievement with cognitive ability. *Intelligence, 35*, 83-92.

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

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab. Acceptance and psychometric characteristics of a computer-based microworld to assess Complex Problem Solving. *Psychological Test and Assessment Modeling, 54*, 54-72.

Van der Linden, W. J. & Glas, C. A. W. (2000). *Computerized adaptove testing.* Dordrecht: Kluwer Academic Publishers.

Veenman, M. V. J. & Elshout, J. J. (1994). Differential effects of instructional support in simulation environments. *Instructional Science, 22*, 363-383.

Wenke, D., Frensch, P. A., & Funke, J. (2005). CPS and intelligence: Empirical relation and causal direction. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 160-187). New York: Cambridge University Press.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex Problem Solving. More than reasoning? *Intelligence, 40*, 1-14.

Table 1

Goodness of Fit Indices for Factor Structure Models (Overall Sample and Subgroups)

Model	χ^2	*df*	*p*	CFI	TLI	RMSEA	n
2-dim. overall sample	178.432	43	.001	.973	.981	.051	1196
1-dim. overall sample	416.071	43	.001	.945	.962	.073	1196
2-dim. only junior high school	43.048	32	>.05	.990	.992	.033	309
2-dim. only senior high school	112.247	39	.001	.940	.951	.062	484
2-dim. only university students	48.210	30	.02	.977	.980	.052	222
2-dim. only blue-collar workers	42.773	18	.001	.966	.963	.087	181

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.

Table 2

Standardized factor loadings and thresholds for models without equality constraints (Model 1) and with equality constraints (Model 2)

Dimension	Item	Model 1: Factor loadings [Thresholds]				Model 2:
		Junior high school	**Senior high school**	**University students**	**Blue-collar workers**	**Common loading [Common Thresholds]**
Knowledge Acquisition	Task 1	.74 [.28]	.62 [-.61]	.77 [-.87]	.70 [.70]	.74* [-.21]*
	Task 2	.77 [-.32]	.62 [-.72]	.78 [-.96]	.71 [-.22]	,73 [-.56]
	Task 3	.91 [-.07]	.91 [-.89]	.79 [-1.10]	.92 [-.04]	,91 [-.53]
	Task 4	.93 [-.47]	.76 [-.69]	.96 [-1.31]	.90 [-.26]	,85 [-.64]
	Task 5	.45 [2.00]	.67 [1.16]	.75 [.80]	.65 [2.13]	,73 [1.3]
	Task 6	.64 [1.42]	.70 [.71]	.81 [.55]	.92 [1.55]	,75 [.92]
Knowledge Application	Task 1	.65 [-.03]	.60 [-.58]	.58 [-.69]	.58 [-.07]	,65 [-.37]
	Task 2	.72 [.83]	.84 [.31]	.87 [.09]	.79 [.45]	,81 [.41]
	Task 3	.88 [.64]	.80 [.13]	.90 [-.01]	.90 [.39]	,85 [.27]
	Task 4	.86 [.16]	.75 [-.13]	.90 [-.63]	.90 [-.05]	,83 [.13]
	Task 5	.59 [.37]	.49 [.12]	.29 [.08]	.52 [.23]	,50 [.21]
	Task 6	.53 [.34]	.42 [.02]	.60 [-.06]	.64 [.33]	,54 [.13]

Note. Model 1: Configural invariance model; Model 2: Common factor loadings and thresholds. Higher thresholds denote more difficult items. * Displays differential item functioning according to Lagrange-Multiplier Tests (χ^2>10.828; df=1; p<.001) All estimates are based on tetrachoric correlations.

Table 3

Latent mean comparisons of knowledge acquisition and knowledge application between subgroups of junior high school students, blue-collar workers, senior high school students, and university students

Dimension	Model	*Compare with*	*M (SE)*	*p*	Cohens' d	*Compare with*	*M (SE)*	*p*	Cohens' d	*Compare with*	*M (SE)*	*p*	Cohens' *d*
Knowledge Acquisition	(1) Junior high school												
	(2) Blue-Collar workers	(1)	-.13 (.10)	>.05	(.10)								
	(3) Senior high school	(1)	.86 (.09)	<.001	.43	(2)	.99 (.11)	<.001	.41				
	(4) University students	(1)	1.16 (.12)	<.001	.65	(2)	1.27 (.13)	<.001	.65	(3)	.35 (.11)	<.001	.21
Knowledge Application	(1) Junior high school												
	(2) Blue-Collar workers	(1)	.14 (.13)	>.05	(.08)								
	(3) Senior high school	(1)	.63 (.13)	<.001	.22	(2)	.43 (.16)	<.001	.12				
	(4) University students	(1)	1.00 (.18)	<.001	.37	(2)	.77 (.20)	<.001	.26	(3)	.27 (.11)	<.01	.16

Note. M = Latent Mean; SE = Standard Error;

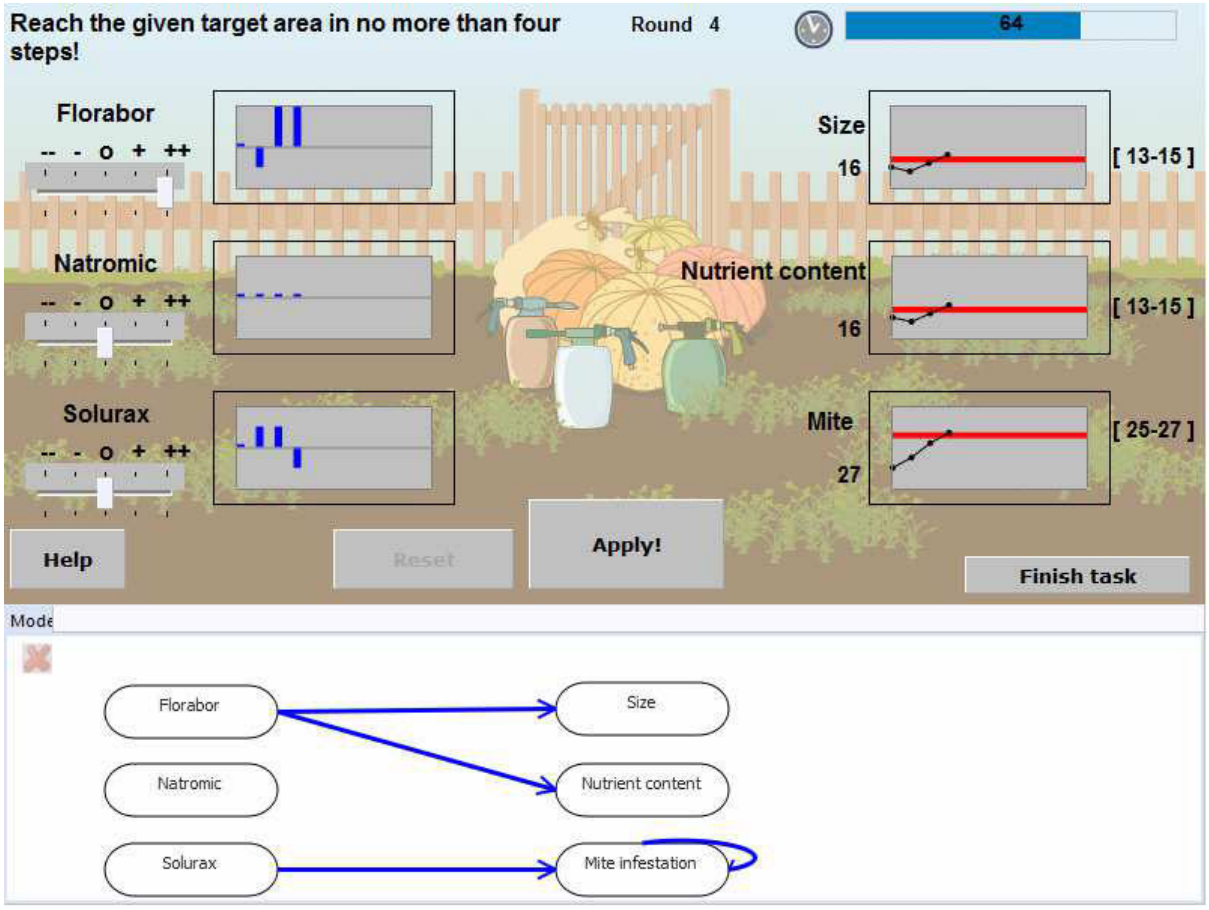

Figure 1. Screenshot of the MicroDYN-task "Planting Pumpkins" at the control phase. The sliders of the input variables range from "- -" (value=−2) to "++" (value=+2). The current value is displayed numerically and the target values of the output variables are displayed graphically and numerically.

Appendix

The 6 tasks were mainly varied regarding two system attributes: the number of effects between variables and the quality of effects (i.e., effects of input and output variables).

Linear Structural Equations	*System size*	*Effects*
$X_{t+1} = 1*X_t + 0*A_t + 2*B_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t$	2x2-System	only effects of inputs
$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2x3-System	only effects of inputs
$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 2*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only effects of inputs
$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only effects of inputs
$X_{t+1} = 1.33*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2x3-System	effects of inputs & outputs
$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_{t+1} = 1.33*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	effects of inputs & outputs

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, whereas X_{t+1}, Y_{t+1}, Z_{t+1} denote the values of the output variables in the subsequent trial.

5

Determinants of Cross-National Gender Differences in Complex Problem Solving Competency.

This article is currently under review:

Wüstenberg, S., Greiff, S., Molnar, G., & Funke, J. (submitted). Determinants of cross-national gender differences in complex problem solving competency. *Learning and Individual Differences.*

Abstract

The present study examined cross-national gender differences in domain-general complex problem solving (CPS) competency and their determinants. A CPS test relying on the MicroDYN approach was applied to a sample of 890 Hungarian and German high school students attending 8th to 11th grade. Results based on multi-group confirmatory factor analyses showed that measurement invariance of CPS was found across gender and nationality. Analyses of latent mean differences revealed that males outperformed females and German students outperformed Hungarian students. However, these results were caused by Hungarian females performing worse than all other groups. Further analyses of logfiles capturing process data of the interaction of participants with the task showed that Hungarian females less often used vary-one-thing-at-a-time (VOTAT) strategy. Results imply that analyzing process data such as use of strategies is highly advisable to identify determinants of performance differences in CPS across groups of interest.

Determinants of cross-national gender differences in complex problem solving competency

1. Introduction

Over the last decades, reports on individual differences in students' performance across gender or nationality have strongly influenced educational policies. For instance, results of the Programme for International Student Assessment (PISA) 2000 lead to changes of the educational system and revisions of educational standards in Germany (Wernstedt & Ohnesorg, 2009). Especially performance differences in domain-specific areas such as mathematical ability play an important role in educational research (Else-Quest, Hyde, & Linn, 2010; Lindberg, Hyde, Petersen, & Linn, 2010), but also in high stakes assessments such as Trends in International Mathematical and Science Study (TIMSS) or PISA.

However, only little is known about individual differences in students' domain-general competencies notwithstanding an increasing scientific and public interest in these competencies. For instance, domain-general problem solving competency will be assessed in PISA 2012 conducted by the Organisation for Economic Co-operation and Development (OECD). More specifically, the OECD emphasizes the high educational and socio-economical relevance of domain-general problem solving in everyday life as it "provides a basis for future learning" (OECD, 2010, p. 7). Thus, domain-general problem solving is considered a highly relevant skill for students that should be developed in addition to domain-specific knowledge within school subjects:

> Mobilisation of prior knowledge is not sufficient to solve novel problems in many everyday situations. Gaps in knowledge must be filled by observation and exploration of the problem situation. This often involves interaction with a new system to discover rules that in turn must be applied to solve the problem (OECD, 2010, p. 15).

Domain-general problem solving, which is referred to as Complex Problem Solving (CPS) in scientific research (Fischer, Greiff, & Funke, 2012; OECD, 2010), includes tasks enabling such interactions between user and task situation (e.g., simulations of technical

devices such as a mobile phone; Wirth & Klieme, 2003). CPS tasks usually contain many highly interrelated elements and tasks' system states change dynamically (cf. Fischer et al., 2012; Funke, 2001). By interacting with CPS tasks, problem solvers have to overcome barriers between a given initial state and a goal state (Funke, 2012; Mayer, 2003). Thereby, they explore and integrate information to discover rules that must be applied to solve the problem (Buchner, 1995). CPS tasks are applied fully computer-based (Wirth & Klieme, 2003), giving researchers the opportunity to not only evaluate outcomes (e.g., if a problem is solved or not), but to analyze process data (e.g., how a problem solver interacts with a problem). This enables analyses of determinants of performance, for instance, which strategies are used to gather information and to solve a certain problem.

While interacting with the task, problem solvers (1) build a problem representation and (2) derive a problem solution (Novick & Bassok, 2005). These two major components of general problem solving are usually measured by two dimensions in CPS research, the competency of problem solvers to gain new knowledge during the interaction with the task - (1) knowledge acquisition - and to apply that knowledge to solve the task - (2) knowledge application (Bühner, Kröner, & Ziegler, 2008; Funke, 2001).

Recently conducted studies show that both dimensions knowledge acquisition and knowledge application can be empirically distinguished (Bühner et al., 2008; Greiff, Wüstenberg, & Funke, 2012; Wüstenberg, Greiff, & Funke, 2012). Furthermore, CPS predicts supervisor ratings (Danner, Hagemann, Schankin, Hager, & Funke, 2011) and school grade point average (Greiff & Fischer, in press; Wüstenberg et al., 2012) even beyond reasoning. However, to our knowledge, no studies have yet been conducted analyzing individual differences in students' CPS performance with regard to gender and nationality.

As a major prerequisite to compare mean performance differences in CPS across gender and nationality, the structure of CPS should not change across groups, that is, structural stability should hold (Byrne & Stewart, 2006; Sass, 2011). Otherwise, between-group

differences could reflect either true differences on CPS, or different psychometric properties of the underlying measurement scale (Brown, 2006). As will be further outlined, it is yet unclear if CPS can be measured with equal validity across (1) gender or (2) nationality, analyses on individual differences in CPS performance are scarce and joint analyses of differences including (3) both gender and nationalities are non-existent. Particularly the latter is of certain interest, because if gender differences vary in specific countries more than in others, cross-national patterns may reflect "inequities in educational and economic opportunities" regarding gender (Else-Quest et al., 2010, p.103).

To this end, based on a sample of Hungarian and German high school students, (1) we will evaluate if CPS can be measured with equal validity across gender and investigate gender differences in mean CPS performance. (2) Further, we will analogously evaluate if CPS can be measured with equal validity across Germans and Hungarians and investigate differences in mean CPS performance. (3) Finally, we conduct combined analyses to study interaction effects of gender and nationality. Therefore, the whole sample is separated in four groups containing German males, German females, Hungarian males, and Hungarian females to evaluate if CPS can be measured with equality validity across gender and nationality and to investigate mean differences across four groups.

1.1 Measurement invariance and latent mean differences across gender

As a prerequisite of interpreting gender differences in CPS, structural stability of the construct has to be secured by evaluating measurement invariance (cf. Byrne & Stewart, 2006), a state of the art procedure frequently applied for measures of cognitive performance (e.g., mathematical ability; Brunner, Krauss, & Kunter, 2008). For instance, it was shown that the factor structure of the Wechsler Intelligence Scale for Children (WISC) does not change across gender (Chen & Zhu, 2008). However, although various CPS measures exist (e.g., *Genetics Lab*, Sonnleitner et al., 2012; *MultiFlux*, Kröner, Plass, & Leutner, 2005; *NewFire*, Rigas, Carling, & Brehmer, 2002; *Tailorshop*, Funke 2010), no studies have been conducted

analyzing measurement invariance with regard to gender. Only recently, it was shown that CPS can be measured invariant across high school students in different grades (Greiff et al., in press).

With regard to gender differences in CPS, previous findings are contrarily to results on reasoning ability, in which reported gender differences slightly favour females showing rather small or marginal effect sizes (Brunner et al., 2008; Halpern & LaMay, 2000; Jensen, 1998). In CPS, only few studies investigated gender differences, pointing towards a considerable advantage of males (Jensen & Brehmer, 2003; Süß, 1996; Wittmann & Hattrup, 2004; Wittmann & Süß, 1999).

For instance, Wittmann and Hattrup (2004) pooled data of three independent studies using the CPS scenario *Tailorshop* (Funke, 2001), in which participants have to maximize the company value of a tailor manufactory by controlling variables such as *number of workers* or *marketing*. In *Tailorshop*, investments in marketing have strong effects on the variable "demand", which in turn increases sales, being highly relevant for good performance within the simulation (Wittmann & Hattrup, 2004, p. 405). The authors showed that males outperformed females (Cohen's $d = .70$) and explained these differences by a higher level of risk aversiveness in females, who invested significantly less in marketing (i.e., varied the variable marketing to a lesser degree) compared to males. However, there are two other possible explanations than less risk aversiveness in females not discussed by Wittmann and Hattrup (2004): (1) Males may rely on more efficient strategies while dealing with CPS tasks or (2) scenario effects may lead to males' better performance.

(1) In cognitive psychology, the use of strategies is known as (implicit) procedural knowledge ("knowing how"), which has to be applied in CPS tasks to identify causal relations between variables that are intransparent to the problem solver at the problem outset (Funke, 2001; Kröner et al., 2005) in order to derive explicit declarative knowledge about the systems' structure ("knowing that"; Kuhn, 2000, p.179). In the study of Wittmann and Hattrup (2004),

males procedural strategy to alter a variable considerably (e.g., making large investments in marketing) is appropriate, because it shows the variables' effect more clearly allowing an easier detection of the systems' causal structure. Even Wittmann and Hattrup (2004) mentioned that "choosing a riskier strategy [creates] a learning environment with greater opportunities to discover and master the rules and boundaries of the game than a more cautious strategy" (p. 406). However, whether males generally use better strategies in CPS tasks has to be proved by applying different CPS tasks besides *Tailorshop* in which other strategies (e.g., vary-one-thing-at-a-time strategy; VOTAT; Tschirgi, 1980) are needed to discover causal relations between variables involved.

(2) Another explanation for the results of Wittmann and Hattrup (2004) is that the environment of a business context in *Tailorshop* may lead to a scenario effect favouring males. For instance, males may be more motivated in "keeping some factory going" as mentioned by Patricia Alexander in a discussion with the authors (cf. Wittmann & Süß, 1999, p.107). Besides such motivational aspects, also prior knowledge about the interplay of marketing demand and sales could have affected performance in *Tailorshop*, as indicated by Süß (1996) who reported that knowledge gathered outside the test situation is significantly correlated with performance in *Tailorshop*. Such scenario effects are criticized by Kröner et al. (2005), who state that CPS tasks should not be influenced "by simulation-specific knowledge acquired under uncontrolled conditions" (p. 349) to assess CPS performance. Just as males might outperform females in *Tailorshop* due to enhanced business knowledge, in an educational context (e.g., testing high school students) males may outperform females in tasks relying strongly on science knowledge due to a better performance in this subject (Kuhn & Holling, 2009; Neuschmidt, Barth, & Hastedt, 2008) and, vice versa, females may outperform males in tasks strongly relying on language (Kuhn & Holling 2009) or verbal memory (Halpern et al., 2007; Kimura, 2002). Thus, the possible effect of motivation or prior knowledge has to be considered, if CPS tasks are embedded in a specific context.

In summary, a part of variance in CPS performance between males and females in the study of Wittmann and Hattrup (2004) might be explained by motivation or prior knowledge. However, effect sizes are large, pointing towards differences in an underlying latent CPS variable, probably caused by males using more efficient strategies. To evaluate if gender differences within the core CPS dimensions *knowledge acquisition* and *knowledge application* hold across different measures, we will use a CPS test based on the MicroDYN approach (Greiff, 2012; Greiff et al., 2012), in which prior knowledge is minimised (cf. method section). We expect that males perform better than females, however, differences should be smaller than in the study of Wittmann and Hattrup (2004) due to the use of tasks in which performance is less influenced by prior knowledge.

Hypothesis 1a: We expect that CPS is measured invariant across gender.

Hypothesis 1b: If measurement invariance is sufficiently met, we expect that latent mean differences between groups indicate significantly better performance of males than females.

1.2 Measurement invariance and latent mean differences across nationality

Measurement invariance of CPS across nationality has not been tested, whereas this has been done for other measures of cognitive performance (e.g., WISC; Chen, Keith, Weiss, Zhu, & Li, 2010; Cognitive Ability Test CogAt; Lakin, 2012). Investigating measurement invariance is necessary when tests are translated from one language to another. Especially, if test items include verbal material such as CPS tasks, their underlying meaning may change during the translation process (Chen, 2008). For instance, items or task descriptions using idiomatic expressions (e.g., item "I feel blue" in a depression questionnaire; Chen, 2008, p. 1006) are not understandable if they are incorrectly translated, leading to non-invariance caused by different patterns of factor loadings or thresholds across groups. Thus, measurement invariance across nationalities has to be tested in order to ensure valid inferences about mean differences.

Equivalently to analyses on measurement invariance, studies dealing with cross-national differences in CPS performance are rather scarce. Strohschneider and Güss (1999) reported that German problem solvers applied more control-oriented strategies than Indian problem solvers. In another cross-cultural study, Güss, Tuason, and Gerhard (2010) used thinking aloud techniques and analyzed verbal protocols to investigate CPS processes across five countries including Germany, Brazil, India, the Phillipines, and the United States. Results based on qualitative indicators showed differences on process and status variables (e.g., amount of information gathered, problem identification, planning, and decision making) showing that problem solving strategies and abilities vary across nationalities.

A comparison of problem solving competency between Hungarians and Germans, however, has only been conducted using paper-and-pencil tasks in PISA 2003 (OECD, 2004). In this large-scale assessment, performance between Hungarian and German students did not differ significantly (OECD, 2004, p. 42). Whilst acknowledging that PISA is not a research venue and that the abilities of actively generating information and using feedback required in CPS is not measured in paper-and-pencil tasks of problem solving (Buchner, 1995; Wüstenberg et al., 2012), we consider these results as a first indicator that students of both countries may not differ in CPS performance. Thus, we expect that Hungarian and German students perform equally within CPS tasks.

Hypothesis 2a: We expect that CPS is measured invariant across nationality.

Hypothesis 2b: If measurement invariance is sufficiently met, we expect no latent mean differences between groups indicating that Hungarians and Germans perform equally well in CPS.

1.3 Cross-national patterns of gender differences

In addition to separate analyses on the relation of gender and nationality with CPS performance, we also analyze interaction effects of both gender and nationality to establish a more detailed picture of CPS abilities and their determinants. Studies investigating such

cross-national patterns of gender differences in problem solving performance have only been conducted using paper-and-pencil tasks as in PISA 2003. Results there show that although in nearly half of the participating countries females outperform male students and vice versa in the other half, these differences are mostly statistically insignificant (OECD, 2004).

Contrarily, in other domains such as Maths or Science, studies report significant interaction effects of gender and nationalities. For instance, a meta analysis based on PISA 2003 and TIMSS 2003 data showed that although mean effect sizes of gender differences in maths are rather small (d<0.15), they differed strongly across countries (ds=-0.42 to 0.40; Else-Quest et al., 2010). Similar results were found for science in TIMSS 2003 (Halpern et al., 2007; Mullis, Martin, Gonzalez, & Chrostowski, 2003) and TIMSS 2007 (Mullis, Martin, & Foy, 2008).

According to Else-Quest et al. (2010), analyzing interaction effects of gender and nationality yield important information on how national characteristics (e.g., "status and welfare of women" or "differences within education systems", p. 125) are related to performance in specific domains. If gender differences vary across Hungary and Germany, this may reflect differences in educational policies in these countries providing important information on education and schooling in the respective countries.

Thus, we analyze interaction effects of gender and nationality by investigating differences in CPS performance using subgroups of German males, German females, Hungarian males, and Hungarian females. However, this analysis is rather exploratory, because although results gathered in PISA 2003 point towards no interaction effects of gender and nationality on problem solving performance (OECD, 2004), it is questionable if these results based on paper-and-pencil tasks can be readily applied to dynamic and interactive measures of CPS, which differ markedly from static paper-and-pencil test of problem solving (OECD, 2010).

Hypothesis 3a: We expect that CPS is measured invariant across gender and nationality, if four groups – German males, German females, Hungarian males, and Hungarian females – are distinguished in the analysis.

Hypothesis 3b: Finally, we expect that analyses of latent mean comparisons between the four subgroups show no interaction effect of gender and nationality. Thus, males are expected to perform better than females in both countries, but effect sizes of performance differences in Germany and Hungary should not vary considerably.

2. Method

2.1 Participants

Data of 890 high school students (433 males) attending 8th to 11th grade were available for analysis. Participation was voluntary and we received signed consent forms from parents of underage students. Participants in the German sample (N =411; 210 males) were recruited from three different school tracks covering all educational levels of the German school system. For the Hungarian sample (N=479; 223 males), we used a subsample of a larger sample and included all participants who attended 8th to 11th grade.[1] Participants in the Hungarian sample were recruited from Hungarian elementary schools (grade 8) and secondary schools (grades 9 to 11).

In the combined German and Hungarian sample, there were nearly as much females in each grade as males and gender distribution neither differed across grade levels 8 to 11 (χ^2=2.00, df=3, p>.05), nor across countries (χ^2=1.83, df=1, p>.05). Missing data due to software problems were excluded on a pairwise basis.

2.2 Material

CPS was measured by a set of tasks based on the MicroDYN approach (Greiff, 2012; Greiff et al., 2012), using several independent problems that rely on the formal framework of linear structural equations (Funke, 2001; see Appendix for equations). Within a MicroDYN

[1] The overall sample conducted in Hungary contained data from students attending grades 5 to 11 and investigated performance differences in CPS across age (Greiff et. al., in press).

task (e.g., the task "perfume" shown in Figure 1), input variables (e.g., fictitious ingredients labelled *Norilan*, *Miral*, *Carumin*; upper left side of Figure 1) influence output variables (e.g., flavours labelled *Fresh*, *Fruity*, *Flowery*; upper right side of Figure 1). Relations between variables are not visible to participants. The procedure within a task is divided into (1) an exploration phase and (2) a control phase.

Please insert Figure 1 about here

In the (1) exploration phase, participants have to identify relations between input and output variables by actively manipulating sliders of the input variables (time frame: 3.5 minutes). For instance, participants may vary solely the value of *Norilan* by pulling a slider from "0" to "++". After clicking on "apply", the value of *Fresh* increases revealing that variables *Norilan* and *Fresh* are related. While exploring, participants represent their conclusions about the relations in a causal diagram (Funke, 2001; see bottom of Figure 1). For instance, participants may draw an arrow between *Norilan* and *Fresh*. By evaluating the correctness of the model drawn, the CPS dimension *knowledge acquisition* is assessed. Subsequently, in the (2) control phase, the correct model is presented to participants and they are asked to reach given target values in each output variable in no more than four steps by manipulating input variables accordingly (time frame: 1.5 minutes). For instance, participants have to increase the value of *Fresh* by setting the slider of *Norilan* or *Miral* on "++". By evaluating if targets are reached, the CPS dimension *knowledge application* is assessed (for a detailed description of the MicroDYN approach see Greiff et al., 2012).

Each task was embedded in a different context to enhance motivation of students (e.g., feeding a cat, training a handball team, mixing chemical elements, producing a perfume, or handling a moped). To avoid uncontrolled influences of prior knowledge, input or output variables were labelled either without deep semantic meaning (e.g., training A, B, and C) or fictitiously (e.g., *Norilan* as a name for an ingredient). Thus, subgroups should not have an

advantage in solving tasks just because of being more familiar with a specific context (e.g., males in "handling a moped" task).

Although the exemplary task “perfume“ contains only effects between input and output variables, a certain output variable may also influence itself (called eigendynamic or autoregressive process) or another output variable (called side effect). Thus, the system state changes either due to participants’ intervention and/or due to dynamics inherent in the system posing additional interactive demands on participants (Funke, 2001; Wüstenberg et al., 2012).

2.3 Dependent variables

Both CPS dimensions, *knowledge acquisition* and *knowledge application*, were scored dichotomously (see Greiff et al., 2012; Kröner et al., 2005). For *knowledge acquisition*, full credit was given if the model drawn by the participants was completely correct and no credit, if participants’ model contained at least one error. For *knowledge application*, full credit was given if target values of all variables were reached and no credit, if at least one target value was not reached.

2.4 Procedure

The CPS test was translated from German to Hungarian by a bilingual translator. In both Germany and Hungary, it was administered at schools’ local computer rooms and lasted about 45 minutes. Afterwards, participants provided demographical data and worked on additional tests that are not discussed in this paper. The CPS test was delivered through the online platform *Testing Assisté par Ordinateur* (TAO; computer based testing) and testing sessions were supervised either by research assistants or by teachers who had been thoroughly trained in test administration.

Testing sessions started with an instruction on how to handle the user interface followed by a trial task. Afterwards, eight MicroDYN tasks were applied to participants. CPS tests used in Germany and Hungary were identical except that the underlying structure of one task differed. This task was not included in all subsequent analyses, although fitting acceptably in

both samples. Furthermore, an additional task was excluded from analyses due to low communality ($r^2_{\text{Hungarian sample}}$=.03; $r^2_{\text{German sample}}$=.07) caused by an extreme item difficulty of *P*=.03 in both samples (*P*=.03). Thus, data analyses were based on six MicroDYN tasks.

2.5 Data Analyses

Within structural equation modeling (SEM; Bollen, 1989), confirmatory factor analyses (CFA) was used to establish a measurement model including the two CPS dimensions *knowledge acquisition* and *knowledge application*, and means and covariance structure (MACS) approach was used to test measurement invariance of CPS and to compare latent means across groups. We applied weighted least squares means and variance adjusted (WLSMV) estimator for categorical outcomes (Muthén & Muthén, 2010) for all analyses, which were conducted in the software package Mplus 5.0 (Muthén & Muthén, 2007).

According to Byrne and Stewart (2006), the first step in testing measurement invariance is to identify a baseline model that fits within the overall sample and in each subgroup. Using this baseline model and in line with our hypotheses, we ran three different measurement invariance analyses testing configural invariance and strong factorial invariance, which were performed identically and varied only in the respective grouping factors gender (Hypothesis 1; male vs. female), nationality (Hypothesis 2; Germans vs. Hungarians), and gender and nationality (Hypothesis 3; German males, German females, Hungarian males, and Hungarian females).

However, the procedure in testing measurement invariance was slightly different from the typical one recommended by Byrne and Stewart (2006), because data on MicroDYN were based on categorical outcomes. Consequently, constraints on model parameters differed in comparison to invariance tests based on continuous outcomes. Thus, in all analyses, we started to test configural invariance by estimating the parameters of the baseline model once again in a multigroup model, in which thresholds and factor loadings are not constrained across groups, factor means are fixed at zero in all groups and residual variances are fixed at

one in all groups as recommended by Muthén and Muthén (2010, p.434). Afterwards, the model of configural invariance was directly compared with the model of strong factorial invariance, in which both factor loadings and thresholds are constrained to be equal across groups, residual variances are fixed at one in one group and free in the other groups, and factor means are fixed at zero in one group (reference group) and free in the other groups (focal groups). Measurement invariance is evaluated by comparing the more restricted strong factorial invariance model with the less restricted configural invariance model in a χ^2-difference test (cf. Byrne & Stewart, 2006; Muthén & Muthén, 2010). If the χ^2- difference test is non-significant, measurement invariance exists. Beyond testing invariance on an overall level, additionally we applied Lagrange-Multiplier tests (LM) to check if a certain group had unwanted advantages on a task level. In LM tests, the global model fit should not significantly increase when specific factor loadings or thresholds of a given task were freed. Otherwise, results indicate that the task does not measure the same construct across groups. In order to analyze latent mean comparisons between groups, we imposed equality constraints on item thresholds and factor loadings and set the latent means of one group - the reference group - to zero (Muthén & Muthén, 2010). Thus, the estimated means for all other groups represent the mean differences in the construct compared to the reference group. Statistical significance of the differences between all groups was determined by z-statistics.

3. Results

3.1 Establishing a baseline model

As a first step to test measurement invariance, a 2-dimensional baseline model including *knowledge acquisition* and *knowledge application* was established within the overall sample and also within each subgroup. According to cut-off values recommended by Hu and Bentler (1999), who suggested that a Comparative Fit Index (CFI) value above .95 and a Root Mean Square Error of Approximation (RMSEA) below .06 indicate a good global model fit, the model showed a good fit in the overall sample (χ^2=129.073, *df*=40, p<.001; CFI=0.975,

RMSEA=0.050; N=890). Both dimensions correlated significantly on a latent level (r=.79). The 2-dimensional model also showed a significantly better fit than a 1-dimensional model (χ^2=229.144, *df*=41, *p*<.001; CFI=0.962, RMSEA=0.072) with *knowledge acquisition* and *knowledge application* combined under one factor as indicated by a significant χ^2-difference test (χ^2=74.489; *df*=1; *p*<.001).

Subsequently, the 2-dimensional model was separately applied to each subgroup – German males, German females, Hungarian males, and Hungarian females. The model fitted well in each group (CFI=0.968 to 0.985, RMSEA=0.022 to 0.054), and, thus, was used as baseline in each of the following analyses (Hypotheses 1 to 3). Communalities for knowledge acquisition (h^2=0.44-0.82) and knowledge application (h^2=0.26-0.79) were mostly above the recommended level of 0.40 (Hair, Anderson, Tatham, & Black, 1998). However, due to the comparable low number of only six administered tasks, internal consistencies of MicroDYN were smaller (*knowledge acquisition* α=.74; *knowledge application* α=.62) than in other studies using MicroDYN tests based on larger item samples (e.g., Greiff et al., 2012; Wüstenberg et al., 2012).

3.2. Hypothesis 1: Gender

Measurement invariance analysis of gender was conducted to determine if the 2-dimensional factor structure of CPS also holds within subgroups of males and females. The fit for the model of strong factorial invariance with factor loadings and thresholds constrained did not differ from the fit of the initial model assuming configural invariance (see first row in Table 1). Thus, CPS is measurement invariant across gender, supporting Hypothesis 1a.

Please insert Table 1 about here

We applied LM-tests to ensure that embedment in a certain context used within a MicroDYN task did not unjustifiably favour a gender group on a specific item level beyond overall differences. This was confirmed, as global model fit did not significantly increase when specific factor loadings or thresholds of any given task were freed.

Regarding latent mean differences across gender, results showed that males performed significantly better in *knowledge acquisition* (M_{Males}=0; $M_{Females}$=-.69, s=.11, p<.001) and *knowledge application* (M_{Males}=0; $M_{Females}$ =-.60, s=.08, p<.001), supporting Hypothesis 1b.

3.3 Hypothesis 2: Nationality

We assumed that CPS is also measured invariant with regard to nationality, implying that translation processes did not affect the construct measured (Chen, 2008). Results showed that measurement invariance held (see second row in Table 1) and LM tests did not yield any significant result, supporting Hypothesis 2a. Concerning mean performance in CPS, we expected that German and Hungarian students did not differ significantly. However, latent mean differences between Germans and Hungarians indicated that Germans performed significantly better in *knowledge acquisition* ($M_{Germans}$=0; $M_{Hungarians}$=-.39, s=.07, p<.001) and *knowledge application* ($M_{Germans}$=0; $M_{Hungarians}$=-.25, s=.10, p<.01). Thus, Hypothesis 2b was not supported.

To summarize results on Hypotheses 1 and 2, CPS was measured invariant across gender as well as nationality and latent mean differences indicated that males outperformed females and German students outperformed Hungarian students.

3.4 Hypothesis 3: Nationality and Gender

In order to allow more elaborated interpretations of the single group result patterns and to analyze cross-national patterns of gender differences, we also checked mean differences of subgroups differentiated by gender and nationality combined (Hypothesis 3). The whole sample was therefore divided into four subgroups: German males (N=210), German females (N=201), Hungarian males (N=223), and Hungarian females (N=256). CPS showed measurement invariance across nationality and gender (see third row in Table 1) and LM tests did not yield any significant result, supporting Hypothesis 3a.

Latent mean differences between German males, German females, Hungarian males and Hungarian females are reported in Table 2. We compared performance between all groups,

starting with the best performing group German males as a first reference group (left column of Table 2).

Please insert Table 2 about here

Results showed that German males performed significantly better in *knowledge acquisition* and *knowledge application* than both Hungarian and German females and were slightly better than Hungarian males, although insignificantly. Subsequently, Hungarian males served as a reference group in a second comparison (middle column of Table 2), outperforming Hungarian females in both CPS dimensions and German females only in *knowledge application*. In a third comparison (right column of Table 2), German females showed a significantly better performance than Hungarian females in both dimensions (overall rank order based on absolute values on CPS performance in both facets: German males > Hungarian males > German females > Hungarian females).

However, statistically significant mean differences between groups do not automatically imply practical relevance and absolute values of latent means can only be interpreted relatively to the reference group in which the mean was fixed, making comparisons between mean scores of *knowledge acquisition* and *knowledge application* inappropriate. For instance, German females had a higher value on *knowledge application* (M=-.43) than on *knowledge acquisition* (M=-.73), but this does not imply that participants performed worse in the latter compared to the former, because the means were not on the same scale.

Consequently, effect sizes were computed to determine practical relevance of results and to allow comparison of performance differences between CPS dimensions (see Table 2). Based on conventions on Cohen's d, an effect size of 0.2 is regarded as small effect, 0.5 as medium effect, and 0.9 as large effect (Cohen, 1988). Accordingly, significant differences between German males, Hungarian males, and German females in both *knowledge acquisition* and *knowledge application* are considered mostly small effects, whereas significant differences between Hungarian females and all other groups showed largely medium effect

sizes (see Table 2). Thus, results indicated that Hungarian females differed markedly from all other groups. This further implies that differences between males and females (Hypothesis 1b) and differences between Germans and Hungarians (Hypothesis 2b) are mostly fostered by low performance of Hungarian females, implying an interaction effect of gender and nationality contrarily to expectations in Hypothesis 3b.

In summary, CPS was measured invariant in all groups. Results on mean differences indicated that males outperformed females and Germans outperformed Hungarians, however, differences mainly resulted from poor performance of Hungarian females. As outlined, neither a change in the construct measured (due to measurement invariance), nor the influence of prior knowledge gathered outside the test situation (as prior knowledge is not helpful to solve the task; cf. method section) sufficiently explained these performance differences. That is, latent mean differences are likely to display real differences in CPS performance. However, the question is how can these differences be explained?

An important indicator of performance in CPS tasks is use of efficient strategies (Vollmeyer, Burns, & Holyoak, 1996), which might have caused superior performance of males in the present study. As mentioned by several researchers, vary-one-thing-at-a-time (VOTAT; Tschirgi, 1980) is an excellent strategy that enables participants to identify isolated effects of one input variable on output variables beyond dynamics of a task (Klahr, 2000; Kuhn, 2000; Vollmeyer et al., 1996). For instance, Wüstenberg et al. (2012) reported that applying VOTAT during the exploration phase in MicroDYN is highly correlated with performance in *knowledge acquisition* and *knowledge application* on a latent level. Within MicroDYN tasks used in this study, VOTAT is an efficient strategy, because only two tasks included effects between output variables (indirect effects; cf. method section), which require the application of additional strategies to detect relations between variables.

3.5 Additional analyses on strategy

To gain deeper insights in potential causes for differences in CPS performance across the four groups, we analyzed logfile data and evaluated use of VOTAT within each group in

MicroDYN. Full credit was given if participants applied VOTAT consistently for all input variables and no credit if VOTAT was used inconsistently or not at all. Results revealed that mean use of VOTAT was highest for German males (M_{votat}=0.76), followed by Hungarian males (M_{votat}=0.65) and German females (M_{votat}=0.63), but was considerably lower for Hungarian females (M_{votat}=0.35), offering a possible explanation for the interaction effect of gender and nationality indicated by notably low performance of Hungarian females in *knowledge acquisition* and *knowledge application* in this study.

4. Discussion

The general aim of this study was to examine, if CPS competency assessed by MicroDYN shows measurement invariance across gender and nationality and to investigate latent mean differences in CPS performance between males and females in a cross-national sample containing Hungarian and German high school students. Analyses further revealed that it is essential to investigate use of efficient strategies in order to yield a detailed picture of determinants of performance in CPS.

4.1 Measurement Invariance

Results provide support that CPS is measured invariant across gender (male vs. female), nationality (Germans vs. Hungarians), and gender and nationality (German males, German females, Hungarian males, and Hungarian females). More specifically, model fit did not deteriorate when factor loadings and thresholds were constrained across groups and LM tests were non-significant yielding three implications: First, latent mean differences can be meaningfully interpreted (Byrne & Stewart, 2006). Second, varying contexts in different CPS tasks (e.g., driving a moped, training a handball team, mixing a perfume), in which influence of prior knowledge is minimised, do not favour a certain group. For instance, males did not outperform females in CPS tasks that are arguably embedded into a male context (e.g., mixing chemical elements) more than in tasks embedded in a female context (e.g., mixing a perfume). Third, results on measurement invariance across nationality show that the process of translating tasks from German into Hungarian language did not affect the construct measured

(Chen, 2008). In summary, performance differences found in our analyses indicate real differences in underlying CPS competency and are unlikely to be a methodological artefact.

4.2 Latent Mean Comparisons

Latent mean comparisons showed that males outperformed females (Hypothesis 1b) and Germans outperformed Hungarians (Hypothesis 2b) in both CPS dimensions. However, cross-national comparisons and further analyses on determinants of performance revealed that results on Hypotheses 1 and 2 were mostly caused by groups' different use of an efficient strategy. That is, females in general and Hungarian females in particular used VOTAT less often than males (see additional analyses on Hypothesis 3b). As mentioned in the introduction, applying strategies such as VOTAT is a prerequisite for gaining declarative knowledge as it is assessed in *knowledge acquisition* (Kröner et al., 2005) and use of strategies is - to a lesser degree - also relevant for solving a problem as it is assessed in *knowledge application* (Wüstenberg et al., 2012), explaining overall performance differences between groups by exploration behaviour.

But why did (especially Hungarian) females apply VOTAT less often? A possible explanation is a missing understanding of the concept of "additive effects [from input variables on output variables] – effects that operate individually on a dependent variable but that are additive in their outcomes" (Kuhn, Black, Keselman, & Kaplan, 2000, p. 498). That means, if two variables A and B have an effect on an outcome variable X (e.g., A positive, B negative), a student who knows that variables have additive effects may more frequently use VOTAT, discovering that effects of both variables cancel each other out. Contrarily, a student who does not understand the principle of additive effects may enhance the amount of both variables only recognizing that the output variable does not change, and, thus, assuming that no variable has an effect (Kuhn et al., 2000). Testing the assumption that a missing understanding of additive effects influenced results was not possible in the present study, but is an interesting venue for further research. For instance, verbal protocols gathered with

thinking-aloud technique may be analyzed to gain more insights in processes involved while participants work on MicroDYN. Nevertheless, missing understanding of additive effects is a reasonable explanation for the worse performance of (especially Hungarian) females.

Assuming that the hypothesis of additive effects may hold, it is still unclear why Hungarian females should be worse in understanding it. During school education, understanding and applying the principle of VOTAT is commonly taught within science education, because VOTAT allows a logical disconfirmation of alternative hypotheses, which is central to most experimental designs (Kuhn et al., 2000; Tschirgi, 1981; Vollmeyer et al., 1996). In the Hungarian school system, science and mathematics are traditionally taught more abstract, pure, and proof oriented compared to international trends (Vári, Tuska, & Krolopp, 2002). Thus, interactive real-world experiments usable to teach domain-general strategies such as VOTAT are less frequently applied. Although this may yield a possible explanation for deficits of Hungarian females, it does not provide an answer on why Hungarian males performed better than Hungarian females. Maybe they compensated lack of knowledge outside school education.

In summary, results showed that performance in *knowledge acquisition* and in *knowledge application* is influenced by use of efficient strategies. Similarly, in the study of Wittmann and Hattrup (2004), males outperformed females because they altered an important variable to a greater extent showing the impact of input variables on output variables more apparently. Although the authors interpreted the result as a consequence of lower risk aversiveness of males, it may also be attributed to applying an appropriate strategy revealing the systems' structure more clearly. Overall, results of Wittmann and Hattrup (2004) as well as our study indicate that males revert to more efficient strategies, which strongly influence performance in CPS. Furthermore, similarly to findings of Else-Quest et al. (2010) in TIMSS 2003 and PISA 2003, gender differences in our study vary considerably across Germans

(*d*s=.20-.25) and Hungarians (*d*s=.38-.43) in both CPS facets, showing a clear interaction effect of gender and nationality.

4.3 Shortcomings and Outlook

MicroDYN rests upon linear structural equation systems, which require a certain degree of mathematical skills, although being kept to a minimum by using only small integer numbers and graphically illustrated target goals (see Figure 1 and Appendix for equations). Thus, other approaches to measure CPS not based on quantitative relations between variables such as finite state automata (Funke, 2001) may be helpful to further investigate whether results depend on the specific operationalization used. Finite state automata tasks rely on qualitative connections between variables and can represent many devices encountered in every day life (e.g., ticket vending machines or mobile phones, Funke & Frensch, 2007; Funke, 2001). For instance, such devices include analogous structures (e.g., menus functioning in a comparable way), which require the application of different strategies compared to tasks within linear structural equation systems to identify a system's causal structure. Thus, comparing CPS performance in linear structural equation and finite state automata tasks may yield additional information on the generalizability of our result based on the VOTAT strategy that males use more efficient strategies in CPS tasks also in general.

Furthermore, to not overburden students, MicroDYN included only two tasks with effects between output variables (indirect effects) and no tasks with interaction effects between input variables in this study. Thus, applying VOTAT is sufficient in most tasks to identify the causal structure of the items. Contrarily, to identify an interaction effect between variables, problem solvers first would have to manipulate two input variables on their own and vary them combined afterwards. For instance, in a MicroDYN task, the output variable *Concrete* may only change if input variables *Water* and *Cement* are indifferent to zero. Enhancing only one input variable (e.g., *Water*) will not have an effect on the amount of *Cement.* Such kind of effects can easily be implemented within the framework of linear

structural equations and would demand other strategies than VOTAT (Kuhn et al., 2000). In doing so, it could be analyzed if males also apply other successful strategies than VOTAT more frequently than females.

Nevertheless, VOTAT is relevant in identifying causal relations between variables in various domains (e.g., biology, economics, physics, psychology) and, therefore, of high importance in educational contexts. In recent years, education seeks to foster students' competencies that are relevant in several domains besides domain-specific knowledge, because it is expected that students transfer what they learned in one domain to other domains (OECD, 2010). In fact, a large amount of education proceeds on the assumption of transfer (Perkins & Salomon, 1989). In this respect, MicroDYN offers great opportunities, because it can be used to teach domain-general strategies such as VOTAT (as well as an understanding of additive effects) or analyses of interaction effects. According to Adam (1989), the more specific the context is, in which thinking skills are trained or knowledge is acquired, the lower the possibility to transfer them to other contexts. Consequently, teaching domain-general strategies in tasks that are embedded in a specific context, for instance science (e.g., Chen & Klahr, 1999; Klahr, Triona, & Williams, 2007) may not be easily transferred to other contexts. Contrarily, MicroDYN tasks can be embedded in various contexts without relying heavily on prior knowledge, allowing an easier transfer to other domains.

However, to enable transfer of knowledge, learners must understand when the application of what has been learned is useful (Bransford, Brown, & Cocking, 1999). Regarding use of strategies, this aspect of meta-cognition is called metastrategic knowledge, which is the ability to know which strategies one has available and to evaluate their usefulness in a specific problem context for reaching a certain goal (Kuhn, 2000). Enhancing metastrategic knowledge is an important developmental and educational goal, because it helps explaining "how and why cognitive development both occurs and fails to occur" (Kuhn, 2000, pp. 178). We suggest using a broad range of CPS tasks (e.g., tasks based on linear structural

equations and finite state automata), requiring different strategies to investigate metastrategic knowledge of students. Analyzing process data gathered during students' interactions with various CPS tasks allows a deeper understanding of cognitive processes engaged.

4.4 Conclusion

In the present study, we investigated cross-national gender differences and their determinants. As it was shown in this study for MicroDYN, an important prerequisite for comparing performance in CPS is to establish a measurement device that allows a meaningful comparison of performance differences (i.e., tasks that are measured invariant). However, analyses on performance differences should not be limited to outcome variables, but focus more on process variables such as use of strategies (e.g., VOTAT) as crucial determinants of differences in final performance of an underlying ability. A sufficient understanding of processes is a requirement of teaching domain-general CPS competencies and enhancing metastrategic knowledge within school education. As Novick and Bassok (2005) stated, "society expects that the problem-solving lessons learned in school - from how to solve math problems to how to design and execute a science fair project to how to analyze literature - will transfer to students' adult lives for the betterment of the world" (p.345). Thus, analyzing domain-general problem solving competency and how it develops will strongly contribute to this goal - an opportunity that should not be missed.

References

Adam, M. J. (1989). Thinking skills curricula: Their promise and progress. *Educational Psychologist, 24*, 25-77.

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

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy Press.

Brown, T. (2006). CFA with equality constraints, multiple groups, and mean structures. In T. Brown (Ed.), *Confirmatory factor analysis for applied research* (pp. 236–319). New York, NY: Guil*df*ord Press.

Brunner, M., Krauss, S., & Kunter, M. (2008). Gender differences in mathematics: Does the story need to be rewritten? *Intelligence, 36*(5), 403-421.

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.* 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.

Byrne, B. M. & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*(2), 287-321.

Chen, H. & Zhu, J. (2008). Factor invariance between genders of the Wechsler Intelligence Scale for Children – Fourth Edition. *Personality and Individual Differences, 45*(3), 260-266.

Chen, H. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.

Chen H., Keith T.Z., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order wisc-iv structure across China, Hongkong, Macau, and Taiwan. *Personality and Individual Differences, 49,* 677–682.

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.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

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.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103-127.

Fischer, A.**,** Greiff, S., & Funke, J. (2012). The Process of Solving Complex Problems. *Journal of Problem Solving, 4*(1), 19-41.

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.

Funke, J. (2012). Complex problem solving. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 682-685). Heidelberg: Springer.

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

Greiff, S. (2012). *Individualdiagnostik der Problemlösefähigkeit*. [Diagnostics of problem solving ability on an individual level]. Münster: Waxmann.

Greiff, S. & Fischer, A. (in press). 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*.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement*, *36*(3), 189-213.

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (in press). Complex Problem Solving in Educational Settings – something beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity. *Journal of Educational Psychology.*

Güss, C. D., Tuason, M. T., & Gerhard, C. (2010). Cross-national comparisons of complex problem-solving strategies in two microworlds. *Cognitive Science, 34,* 489-520.

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

Halpern, D. F., Benbow, C., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8,* 1-51.

Halpern, D. F. & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review, 12,* 229−246.

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.

Jensen, A. R. (1998). *The g factor.* The science of mental ability. Westport: Praeger.

Jensen, E. & Brehmer, B. (2003). Understanding and control of a simple dynamic system. *System Dynamics Review, 19*(2), 119-137.

Kimura, D. (2002). Sex hormones influence human cognitive pattern. *Neuroendocrinology Letters Special Issue Supplement 4, 23,* 67-77.

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

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

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

Kuhn, D. (2000). Metacognitive development. *Current directions in Psychological Science, 9*(5), 178-181.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The Development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*(4), 495-523.

Kuhn, J. T. & Holling, H. (2009). Gender, reasoning ability, and scholastic achievement: A multilevel mediation analysis. *Learning and Individual Differences, 19*(2), 229-233.

Lakin, J. M. (2012). Multidimensional ability tests and culturally and linguistically diverse students: Evidence of measurement invariance. *Learning and Individual Differences, 22*(3), 397-403.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New Trends in Gender and Mathematics Performance: A Meta-Analysis. *Psychological Bulletin, 136*(6), 1123–1135.

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

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2003). *Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth*

Grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

Muthén, L. K. & Muthén, B. O. (2010). *Mplus User's Guide* (6th edition). Los Angeles, CA: Muthén & Muthén.

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995-2003). *Studies in Educational Evaluation, 34*(2), 56-72.

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

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

OECD (2010). *PISA 2012 Problem Solving Framework.* Paris: OECD Publishing.

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

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

Sass (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347-363.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab. Acceptance and psychometric characteristics of a computer-based microworld to assess Complex Problem Solving. *Psychological Test and Assessment Modeling, 54* (1), 54-72.

Strohschneider, S. & Güss, D. (1999). The Fate of the Moros: A Cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology, 34*(4), 235-252.

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.

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

Vári, P., Tuska, A., & Krolopp, J. (2002). Change of emphasis in the Mathematics assessment in Hungary. *Educational Research and Evaluation, 8*(1), 109-127.

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.

Wernstedt, R. & John-Ohnesorg, M. (2009). *Bildungsstandards als Instrument schulischer Qualitätsentwicklung* [Educational standards as an instrument of quality management]. Bonn: Bonner Universitäts-Buchdruckerei.

Wirth, J. & Klieme, E. (2003). Computer-based Assessment of Problem Solving Competence. *Assessment in Education: Principles, Policy & Practice, 10(3),* 329-345.

Wittmann, W. & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393-440.

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.

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

Table 1

Goodness of Fit Indices for Measurement Invariance of CPS

Group	Invariance model	χ^2	*df*	*p*	CFI	TLI	RMSEA	*free Par.*	*Compare with*	$\Delta\chi^{2\,(1)}$	$\Delta df^{(1)}$	p
gender	(1) Configural Invariance	141.029	68	<.001	.970	.980	.049	50				
	(2) Strong Factorial Invariance	139.415	73	<.001	.980	.984	.045	42	(1)	2.557	7	>.10
nationality	(1) Configural Invariance	134.934	71	<.001	.980	.984	.045	50				
	(2) Strong factorial invariance	132.696	76	<.001	.983	.987	.041	42	(1)	1.545	7	>.10
nationality & gender	(1) Configural Invariance	158.732	107	<.001	.983	.984	.047	100				
	(2) Strong Factorial Invariance	159.635	116	<.01	.986	.988	.041	76	(1)	11.420	16	>.10

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; HF = Hungarian females; HM = Hungarian males; GM = German males; GF = German females.

[1)] χ^2 and Δdf were estimated by a χ^2-difference test procedure in MPlus (see Muthén & Muthén, 2010); χ^2-differences cannot be compared by subtracting χ^2 and *df* if WLSMV-estimators are used.

Table 2

Latent Mean Comparisons of Knowledge Acquisition and Knowledge Application Between Nationality and Gender

Dimension	Model	*Compare with*	*M (SE)*	p	Cohens *d*	*Compare with*	*M (SE)*	p	Cohens *d*	*Compare with*	*M (SE)*	p	Cohens *d*
Acquisition	(1) German males												
	(2) Hungarian males	(1)	-.23 (.12)	>.05	(.13)								
	(3) German females	(1)	-.74 (.25)	<.001	.20	(2)	-.33 (.20)	>.05	(.11)				
	(4) Hungarian females	(1)	-.86 (.10)	<.001	.54	(2)	-.69 (.10)	<.001	.43	(3)	-.55 (.09)	<.001	.38
Application	(1) German males												
	(2) Hungarian males	(1)	-.11 (.13)	>.05	(.06)								
	(3) German females	(1)	-.43 (.12)	<.001	.25	(2)	-.36 (.12)	<.01	.21				
	(4) Hungarian females	(1)	-.81 (.12)	<.001	.42	(2)	-.73 (.12)	<.001	.38	(3)	-.38 (.17)	<.05	.14

Note. M = Latent Mean; SE = Standard Error.

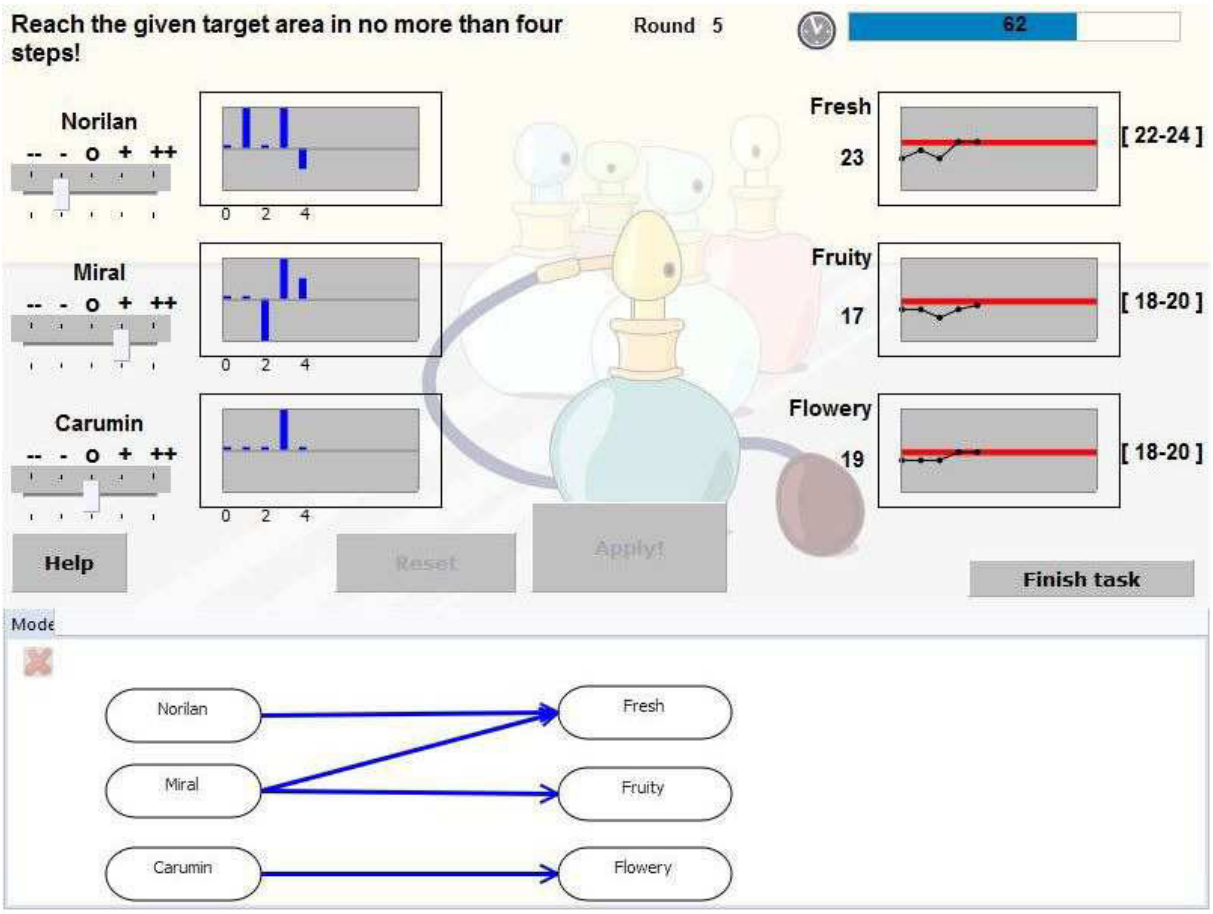

Figure 1. Screenshot of the MicroDYN-task "perfume" during control phase. The sliders of the input variables range from "- -" (value=–2) to "++" (value=+2). Current values and target values are displayed graphically and numerically.

Appendix

The six tasks in this study were mainly varied with regard to two system attributes proved to be most influential on difficulty (see Greiff, 2012): the number of effects between variables and the quality of effects (i.e., effects of input and output variables).

	Linear Structural Equations	***System size***	***Effects***
Task 1	$X_{t+1} = 1*X_t + 0*A_t + 2*B_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t$	2x2-System	only effects of inputs
Task 2	$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2x3-System	only effects of inputs
Task 3	$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 2*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only effects of inputs
Task 4	$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	only effects of inputs
Task 5	$X_{t+1} = 1.33*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2x3-System	effects of inputs & outputs
Task 6	$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_{t+1} = 1.33*Z_t + 0*A_t + 0*B_t + 2*C_t$	3x3-System	effects of inputs & outputs

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, whereas X_{t+1}, Y_{t+1}, Z_{t+1} denote the values of the output variables in the subsequent trial.

6

General Discussion

6. Discussion

This thesis aimed at enhancing the understanding of the nature and validity of CPS. The focus was on extending previous research on CPS by addressing several issues that have not been sufficiently regarded beforehand: For the first time, (1) the internal structure of CPS comprising the two dimensions *knowledge acquisition* and *knowledge application* was investigated across samples with varying cognitive performance and (2) structural stability of CPS was tested across samples varying in gender, age, and nationality. Furthermore, performance differences across these groups were identified and possible explanations were provided. Finally, (3) construct validity of CPS was analysed by relating performance in CPS to a measure of general mental ability. To test hypotheses on these issues, an existing version of MicroDYN was adapted to measure a broad range of cognitive performance.

In summary, concerning (1) internal structure, a 2-dimensional model fitted best and results on (2) structural stability showed that the construct CPS was measured equally well across all samples. Thus, performance differences in MicroDYN tasks occurred due to real differences in CPS. As expected, participants performed better the higher they were educated. However, further analyses revealed that the vary-one-thing-at-a-time (VOTAT) strategy partly explained performance differences across groups. Regarding (3) construct validity, CPS showed incremental validity beyond measures of general mental ability.

In this chapter, results gathered from all papers on the three research questions concerning internal structure, structural stability, and construct validity are tied together and further elaborated by adding aspects that have not been addressed in Paper 1 to Paper 4. Afterwards, strengths of the present research are mentioned and limitations combined with directions for future research are pointed out, followed by a general conclusion.

6.1.1 Internal structure of CPS

In all papers that are part of this thesis, 2-dimensional models of CPS containing the dimensions *knowledge acquisition* and *knowledge application* showed exceptionally good fit indices for overall samples (CFI=.94-.99; RMSEA=.02-.06) and fitted significantly better than 1-dimensional models (CFI=.91-.98; RMSEA=.06-.08) according to χ^2- difference tests. For the first time, the theoretically defined dimensions of *knowledge acquisition* and *knowledge application* could be empirically distinguished using samples with a broad range of cognitive performance including junior high school students as well as blue-collar workers in addition to university students and senior high school students.

Besides the two main dimensions of CPS, *knowledge acquisition* and *knowledge application* (cf. Funke, 2001), Kröner, Plass, and Leutner (2005) and Greiff, Wüstenberg, and Funke (2012) showed that *use of strategies* as prerequisite for gaining knowledge can be modelled as third dimension. *Use of strategies* is commonly operationalized by application of VOTAT, which is an efficient strategy to identify causal relations within a set of variables (Kuhn, 2000; Kröner et al., 2005; Greiff et al., 2012; Vollmeyer, Burns, & Holyoak, 1996). However, in Paper 1, the 3-dimensional model with *use of strategies* as third dimension (operationalized by scoring use of VOTAT) did not fit significantly better than a 2-dimensional model with an aggregated dimension of *use of strategies* and *knowledge acquisition* on one factor and *knowledge application* on another factor. Furthermore, *use of strategies* and *knowledge acquisition* were highly correlated on a latent level (r=.97, p>.001). Although only fit indices of well-fitting 2-dimensional models were reported in Paper 2 to Paper 4, subsequent analyses showed that within these samples *use of strategies* and *knowledge acquisition* were also highly correlated on a latent level (Paper 2: r= 96, p<.001; Paper 3: r= 99, p<.001; Paper 4: r= 97, p<.001), indicating that 3-dimensional models did not provide additional information on CPS performance.

However, why were these results contrary to findings of Kröner et al. (2005) and to the first version of MicroDYN (Greiff, 2012; Greiff et al., 2012), in which three dimensions could be empirically distinguished and correlations between *use of strategies* and *knowledge acquisition* have been considerably smaller? Differences to Kröner et al. (2005) might be allocated to the operationalization of *knowledge acquisition*. Kröner et al. (2005) did not only assess knowledge about the causal structure in *knowledge acquisition*, but also partly *knowledge application* competency (cf. Paper 1). Varying results in contrast to the first MicroDYN version might be at least partly due to the changes carried out to reduce difficulty of tasks and to minimize influence of mathematical ability.

In the current version of MicroDYN applied in this thesis, path weights of variable relations did not vary considerably across tasks and participants only had to draw relations between variables in a causal model instead of both drawing relations and specifying path weights (cf. section 1.3.1). VOTAT, as indicator for *use of strategies*, is a suitable strategy for identifying paths between variables, whereas basic mathematical ability helps to identify path weights that determine how strong input and output variables are related. As path weights are not included in the score of *knowledge acquisition* in the current version of MicroDYN compared to the first version, basic mathematical knowledge was not that much needed in addition to VOTAT for *knowledge acquisition* scores. This led to an increased correlation between *use of strategies* and *knowledge acquisition* in the studies included in this thesis. Furthermore, fewer tasks with indirect effects were applied to participants in the current MicroDYN version, in which VOTAT is not as efficient as in tasks with only direct effects to detect causal relations (Greiff, 2012; Greiff et al., 2012). This may also have increased correlations between *use of strategies* and *knowledge acquisition*.

In summary, the changes made to minimize influence of mathematical ability and to decrease difficulty of the tasks increased the influence of VOTAT on performance in

knowledge acquisition compared to the first MicroDYN version with varying path weights across tasks and many indirect effects. This led to the acceptance of the 2-dimensional model in all papers included in this thesis. In general, a 2-dimensional structure will most likely result empirically if strategies used to identify causal relations in a given task can be completely identified and measured. The more strategies necessary for appropriate *knowledge acquisition* are *not* considered in the measurement of *use of strategies*, the higher the possibility of yielding a 3-dimensional model of CPS.

6.1.2 Structural stability and performance differences in CPS

Paper 2 to Paper 4 investigated structural stability of CPS by evaluating measurement invariance of MicroDYN. In detail, measurement invariance held across high school students of different ages (Paper 2), across samples varying considerably in age and educational background (Paper 3), and across gender and nationality (Paper 4). That is, constraining factor structure of CPS, factor loadings of items, and item difficulties to be identical across groups did not worsen the model fit. This implied structural stability of the construct, allowing a valid interpretation of mean score differences across groups (Byrne & Stewart, 2005).

For instance, regarding gender, identical item difficulties for males and females indicated that having different coverstories embedded in male context (e.g., mixing chemical elements) did not unjustifiably reward males. Accordingly, tasks that were arguably embedded in a female context (e.g., mixing a perfume) did not unjustifiably reward females. Thus, applying different coverstories did not influence the performance of specific groups more than others. However, due to written feedback provided by participants after tasks were administered, varying coverstories enhanced participants' motivation while working on MicroDYN.

Similarly, translating the tasks from German into Hungarian did not change the structure of the construct. Consequently, MicroDYN is suitable for application in foreign

countries, although structural stability has still to be proven in a cross-cultural design (e.g., including countries that differ more strongly in the way of life than Germany and Hungary). In summary, MicroDYN is the only measurement device of CPS, in which structural stability across groups was evaluated, implying that performance differences in MicroDYN can be allocated to real differences in CPS.

Analyses of latent mean differences showed that older high school students outperformed younger ones in both *knowledge acquisition* and *knowledge application* with exception of students in Grade 9 (cf. Paper 2). Furthermore, participants with higher educational background did significantly better than participants with lower educational background (cf. Paper 3). Finally, Germans outperformed Hungarians and males outperformed females (cf. Paper 4).[1]

However, subsequent analyses revealed that performance differences in *knowledge acquisition* and, to a lesser extent, in *knowledge application*, could at least be partly explained by *use of strategies*. In more detail, the VOTAT strategy was highly correlated with *knowledge acquisition* (e.g., Paper 1) and considerably correlated with *knowledge application*. If the proportion of participants who applied VOTAT in a given task is descriptively compared with the proportion of participants who drew the correct model in the respective task, the following pattern is revealed (cf. Table 1, also see Paper 1): In easy tasks regarding *knowledge acquisition* (i.e., high number of correct models), more participants drew the correct model than applied VOTAT, indicating that VOTAT is not necessarily needed for identifying relations. In tasks with medium difficulty (i.e., medium number of correct models), the frequency of applying VOTAT and having the correct model drawn was nearly identical, indicating that VOTAT seemed to be very helpful for identifying relations in these tasks. In difficult tasks containing indirect effects (i.e., low number of correct models), the frequency

[1] This result would not change, if Grade 9 in Hungary and Germany was excluded from all analyses. Grade 9 in the Hungarian sample showed a significantly worse performance than all other grades, eventually due to motivational effects, and was therefore excluded from analyses on construct validity in Paper 2.

of applying VOTAT was much higher than the frequency of correct models, indicating that VOTAT was not sufficient for identifying relations in these tasks. However, only few tasks with indirect effects were used in the current version of MicroDYN, explaining the overall high correlation of *use of strategies* measured by VOTAT and *knowledge acquisition*. The descriptive finding is also supported by item-based correlations between *use of strategies* and *knowledge acquisition*, showing high correlations between both dimensions in Item 3 to 5 (r=.80-.85; p<.01) and smaller correlations in Items 1 to 2 (r=.41-.62; p<.01) and Items 6 to 8 (r=.03-.21; p=.01-.64). In summary, usefulness of VOTAT as a strategy depended on the task characteristics and did not automatically lead to better *knowledge acquisition* scores.

Table 1

Proportion of Participants in Paper 1 Who Applied VOTAT (Use of Strategies), Drew the Correct Causal Model (Knowledge Acquisition,) and Reached Target Goals (Knowledge Application)

Item	Task Characteristics	Use of Strategies		Knowledge Acquisition		Knowledge Application
Item1	only direct effects; 2 relations	0.74	<	0.81	>	0.76
Item2	only direct effects; 3 relations	0.77	<	0.83	>	0.47
Item3	only direct effects; 4 relations	0.84	≈	0.83	>	0.62
Item4	only direct effects; 4 relations	0.87	≈	0.86	>	0.50
Item5	only direct effects; 4 relations	0.90	≈	0.90	>	0.74
Item6	direct & indirect effects; 3 relations	0.89	>	0.21	<	0.47
Item7	direct & indirect effects; 3 relations	0.90	>	0.29	<	0.52
Item8	direct & indirect effects; 4 relations	0.90	>	0.07	<	0.70

Note. Use of strategies, *knowledge acquisition*, and *knowledge application* were scored dichotomously. For *use of strategies*, full credit was given if participants consistently applied VOTAT for all input variables, otherwise no credit was scored. For *knowledge acquisition*, full credit was given if the model drawn by the participants was completely correct and no credit, if participants' models contained at least one error. For *knowledge application,* full credit was given if all target goals were reached and no credit, if at least one goal was not reached.

A comparison of item difficulties in *knowledge acquisition* and *knowledge application* across groups revealed that tasks, in which *knowledge acquisition* was rather easy, were not necessarily easy tasks to control, and vice versa. Nevertheless, latent mean differences in *knowledge acquisition* across groups were comparable to performance differences in *knowledge application* across these groups (cf. Paper 2 to Paper 4), showing a convergent pattern of performance differences in these CPS dimensions. Put differently, participants who outperformed others in *knowledge acquisition* tended to show better performance in *knowledge application*, too.

6.1.3 Construct validity of CPS

Paper 1 and Paper 2 investigated construct validity of CPS. The main focus was on the relation of CPS and measures of general mental ability. CPS was correlated with reasoning, however, significant proportions of variance remained unexplained (Paper 1). Even after controlling for reasoning, CPS explained additional variance in school grades incrementally beyond reasoning in a university student sample (Paper 1) and, to a lesser extent, in a sample of high school students (Paper 2). Furthermore, in another study not included in this thesis, it was shown that MicroDYN also explained variance in school grades beyond measures of working memory (Schweizer, Wüstenberg, & Greiff, in press).

These fundamental results indicate that CPS may assess a competency not captured in traditional measures of general mental ability. CPS includes dynamic and interactive aspects of generating and applying knowledge that may constitute a part of general mental ability that is not captured yet in Carroll's three stratum theory of intelligence (Carroll, 2003; cf. Danner, Hagemann, Schankin, Hager, & Funke, 2011). In this theory, a latent factor of general mental ability is located at the third stratum, which explains variance of correlated factors on the second stratum. On the first stratum there are specific abilities that are considered as subcomponents of the factors on the second stratum. As suggested in Paper 1, CPS may be

part of the second stratum and the dimensions of *knowledge acquisition* and *knowledge application* may be located at the first stratum, however, both assumptions have still to be proven (cf. section 6.3).

Another explanation for the unshared variance of CPS and reasoning is that computer knowledge may affect CPS, but not reasoning. Unlike reasoning tasks, CPS tasks require a lot of interaction with computers, implying that more computer knowledge may be necessary to solve a MicroDYN task than a computerized version of a reasoning task. Going one step further, computer knowledge may also explain performance differences in CPS between males and females, because the former arguably play more computer games that are similar to MicroDYN tasks (Cherney & London, 2006). In the study of Paper 1, procedural computer knowledge was assessed by INCOBI-R questionnaire (Richter, Naumann, & Horz, 2010), but was not implemented as variable in the paper in order not to overstretch the length of the manuscript. However, further analyses revealed that standardized path coefficients of procedural computer knowledge predicting performance in CPS (*knowledge acquisition*: β=.46, p<.001; *knowledge application*: β=.42, p<.001) and in reasoning (Advanced Progressive Matrices: β=.47, p<.001) were virtually identical. Furthermore, the increment of CPS beyond reasoning in explaining variance in school grades persisted even if procedural computer knowledge was partialled out (χ^2=130.487, p<.05; CFI=.968; RMSEA=.035). These results imply that the influence of procedural computer knowledge on performance in CPS was comparable to the influence on computer-adapted reasoning tasks, although the CPS test MicroDYN requires considerably more interaction with the computer.

Finally, as shown in Paper 2, parental education was significantly related to CPS in earlier grades. This may imply that CPS is influenced by environmental preconditions. For instance, more highly educated parents may offer more appropriate learning environments consisting of interactive and dynamic situations, which enhance children's ability to acquire

and apply knowledge. However, this tentative result on the relation of CPS and parental education has to be interpreted in a cautious way.

In summary, this thesis replicated well-established results of previous research showing that CPS and traditional measures of general mental ability are significantly related (e.g., Kröner et al., 2005; Leutner, 2002). However, it goes beyond existing findings by demonstrating that a psychometrically adequate assessment of CPS results in incremental validity beyond traditional measures of general mental ability. Research on the influence of parental education on CPS has just begun and further studies are required to allow more valid interpretations about preconditions of CPS.

6.2 Strengths

Results discussed were based on extraordinary large sample sizes compared to previous research on CPS with *N* exceeding 150 in each study and a total of *N*= 2051 participants. Having such large data sets made it possible to run state-of-the-art analyses using confirmatory factor analyses and structural equation modelling, which requires a sufficient number of data points (Bollen, 1989). Samples were based on high school students, university students, and adults, resulting in increased generalizability of results. Further, papers in this thesis combined psychometrical research on CPS with theoretically driven analyses of the reasons why performance outcomes occurred the way they did.

Concerning psychometrics, it was shown for the first time that the internal structure of CPS also held in non-university samples, that CPS was measured invariant across groups varying in age, gender, and nationality, and that CPS showed incremental validity beyond traditional measures of general mental ability. Furthermore, this thesis contributed to a theoretical understanding of CPS by relating results on internal structure of CPS to concrete operationalizations of the LSE-tasks used. In-depth analyses of the interplay of the dimensions *use of strategies*, *knowledge acquisition*, and *knowledge application* were

provided, supplying important information on the influence of adequate *use of strategies* for performance in the latter two dimensions. In this respect, it cannot be emphasized strongly enough that analysing process data such as *use of strategies* is the most promising approach to identify underlying causes of performance differences, as will be outlined in section 6.3.

6.3 Limitations and outlook

Limitations and open questions of the present work pertain to the following aspects that are subsequently addressed. (1) By restricting task variety in MicroDYN, influence of VOTAT on CPS performance is enhanced, necessitating a broader operationalization of CPS. (2) Analyses on convergent and criterion validity of MicroDYN have to be extended. (3) Research on how CPS can be integrated in Carroll's three stratum theory of intelligence (Carroll, 1993) should be widened by applying multiple measures of general mental ability and CPS. (4) Further options to study process data are recommended. These should go beyond analyses in Paper 4 that have already yielded informative results on why performance outcomes in CPS occurred. (5) Finally, longitudinal studies that also deal with the trainability of CPS competency may give further information on how CPS competency develops.

(1) *Influence of VOTAT on MicroDYN performance*: As outlined above (cf. section 6.1.2), in most tasks applied in studies in this thesis the use of VOTAT was significantly related to *knowledge acquisition*. In general, VOTAT and its inherent principle of isolated variation of variables is an efficient strategy applied to identify causal relations between variables in many application areas (Chi & vanLehn, 2007; Chen & Klahr, 1999; Kuhn, 2000; Vollmeyer et al., 1996). Thus, it is a domain-general strategy that should be included in the measurement of CPS. However, to widen the operationalization of the CPS construct measured and to increase external validity, MicroDYN tasks should be constructed in a way that different strategic approaches are needed to acquire knowledge.

One opportunity to extent MicroDYN is to integrate an “interaction effect”, which requires participants to apply other strategies than VOTAT. Within the framework of LSE-tasks, an interaction effect implies that two input variables will have an effect on an output variable if both input variables are varied in conjunction (cf. Paper 4). Such effects can be frequently found in complex problems in daily life. For instance, in a manufactory, increasing the amount of machines will increase productivity of a given product just as buying more raw materials will. However, beyond a certain point, increasing only the number of machines will not enhance productivity, because machines stand still without enough raw materials available. Accordingly, buying more raw materials without increasing the number of machines will not lead to higher productivity, because the production capacity of machines is limited.

(2) *Convergent and criterion validity:* In this thesis, no CPS test in addition to MicroDYN was applied to check convergent validity of MicroDYN. Although previous results based on the initial MicroDYN version showed that performance in *Space Shuttle* and MicroDYN were significantly correlated (Greiff et al., 2012), further research is needed to test if MicroDYN and other measures of CPS assess the same construct. A more sophisticated approach to analyse convergent validity that goes beyond mere correlational studies is to apply latent state trait (LST) theory (Danner et al., 2011; Steyer, Schmitt, & Eid, 1999) or a Multi-Trait-Multi-Method (MTMM) approach (Campbell & Fiske, 1959).

In LST theory, a given trait is measured on two occasions by at least two different measurement devices. The measurement of a given variable can be decomposed into parts dedicated to a given trait, a state residual, a method residual, and an unsystematic measurement error (Steyer et al., 1999). A trait is consistently measured, if several devices assess the same aspects across different occassions, indicating high trait specificity, low method specificity, and low state specificity. Danner et al. (2011) used LST theory to show that the latent CPS factor consisting of performance in *Space Shuttle* (Wirth & Funke, 2005)

and *Tailorshop* (Funke, 2001; Süß, 1996) revealed significant but rather small trait specificities. Thus, in their experiment, common variance of devices was rather small, indicating that they partly measured different aspects. Performance in *Tailorshop* (cf. section 1.2) and *Space Shuttle* depend on both CPS competency and prior knowledge. In *Space Shuttle*, for instance, the landing gear has to be retracted when leaving a planet with the shuttle (Wirth & Funke, 2005). This action is familiar to participants who have knowledge about space flights or flights in general. Thus, the assessment of different kinds of prior knowledge in *Space Shuttle* and *Tailorshop* may lead to small trait specificities as mentioned by Danner et al. (2011). However, even with small trait specificity, the latent CPS factor explained variance in supervisory ratings beyond measures of reasoning, indicating the potential of CPS (Danner et al., 2011).

As further development, trait specificity of CPS measures may be enhanced if approaches are used that both assess genuine aspects of CPS not confounded with prior knowledge such as MicroDYN and MicroFIN. MicroFIN, which is currently developed at the University of Heidelberg, uses multiple tasks based on finite state automata to measure CPS competency (Greiff et al., 2010). Because CPS performance in MicroDYN and MicroFIN is not confounded with specific prior knowledge about content due to variable labels without deep semantic meaning, trait specificity of these two measurement devices of CPS is expected to be larger than in devices, in which different prior knowledge affects performance.

Convergent validity could also be tested by applying a MTMM design differing slightly from LST, in which two or more traits are measured on only one occasion (Eid, Lischetzke, & Nussbeck, 2006). This allows a combined testing of multiple CPS dimensions (e.g., *knowledge acquisition*, *knowledge application, use of strategies*) measured by several CPS tasks. If common trait variance across devices is high and method-specific influences and

unsystematic measurement errors are low, the different CPS tests assess the same construct supporting convergent validity of MicroDYN

Finally, concerning criterion validity, performance in MicroDYN explained significant amounts of variance in school performance. However, it should also be significantly related to other criteria involving problem solving performance in real life. Recently, it was shown that MicroDYN explained performance in technical problem solving in a sample of electronics technicians (Abele et al., 2012). In this study, participants' technical problem solving competency was measured by their ability to identify malfunctions in a computer-simulation of electrotechnical systems. Another approach to prove criterion validity would be to test if performance in MicroDYN distinguishes between advanced and less advanced problem solvers in regular job situations. However, identifying good problem solvers at work with accurate measurement devices is quite challenging.

(3) *CPS and general mental ability*: This thesis showed that CPS measured by MicroDYN is incrementally valid beyond a measure of general mental ability. However, comparable to other research (e.g., Danner et al., 2011), reasoning tests (i.e., Advanced Progressive Matrices, Raven, 1958; Culture Fair Test 20R, Weiß, 2006) or working memory (Schweizer et al., in press) were chosen as indicators of general mental ability. Although reasoning is seen to be at the core of general mental ability (Carroll, 1993), widening the operationalisation would reveal more valid results on how CPS can be integrated in Carroll's (1993) three stratum theory of intelligence. In this respect, measures of general mental ability should include a test battery comprising factors on the second stratum such as cognitive speediness, processing speed, visual perception, or crystallized intelligence. However, to the best knowledge of the author, "self-regulated learning and the ability to adapt the problem solving process to a changing environment by continuously processing feedback" (Wirth &

Klieme, 2003, p. 329) as captured in CPS is not included in factors at the second stratum, making assessment of this cognitive aspects unique to CPS.

Nevertheless, to ensure that incremental validity of CPS found in this thesis does not only relate to the specific configuration of the tasks used, it has to be shown that common aspects of two or more measures of CPS are incrementally valid beyond two or more traditional measures of general mental ability. Therefore, the MTMM approach already described may also be used to yield results on incremental validity of CPS. If common trait variance of several CPS measures explains performance in an important performance criterion beyond common trait variance of multiple measures of established intelligence tests, usefulness of CPS tests in addition to traditional measures of general mental ability is more evident.

(4) *Process data*: Paper 4 in this thesis emphasized the importance of analysing process data in order to get a clearer picture of what determines performance in CPS competency. On the one hand, process measures can be gathered by using verbal protocols that allow researchers to investigate misconceptions of participants (e.g., Swanson, 1990; Kuhn, 2011), either by applying thinking-aloud techniques while participants work on a CPS task or by asking participants after test administration why they responded to a task the way they did (Kuhn, Black, Keselman, & Kaplan, 2000).

On the other hand, the CPS dimension *use of strategies* can be directly extracted from logfiles recording participants' interactions with the task environment as realized in Paper 1 and Paper 4. As research on problem solving shows, appropriate *use of strategies* while exploring the problem space of a given task results in good performance in that task (e.g., Kröner et al., 2005; Paper 1 and Paper 4 in this thesis). However, is there an underlying competency that determines performance in *use of strategies* across several tasks requiring different strategies?

Metastrategic knowledge may explain performance in *use of strategies* across tasks, that is, a part of metacognition representing the ability to know which strategies are available to a person and to evaluate their usefulness in a specific problem context for reaching a certain goal (Kuhn, 2000). For instance, Swanson (1990) showed that regardless of their cognitive ability, students with higher metastrategic knowledge (i.e., metacognition measured by *use of strategies*) outperformed students with lower metastrategic knowledge in a hands-on problem solving task. However, Swanson's experiment as well as studies in this thesis did not assess common variance of different strategies across diverse tasks. In this respect, MTMM models may allow a more general evaluation of how CPS performance is influenced by metastrategic knowledge. Therefore, the CPS dimension *use of strategies* has to be measured in several CPS tests (e.g., MicroDYN, *MicroFIN*, *Tailorshop*). As this thesis shows, appropriateness of strategies is dependent on the tasks used due to varying task constraints (cf. section 1.4.3), implying that efficient strategies in CPS tasks such as *Tailorshop* may considerably differ from efficient strategies in MicroDYN. For instance, if participants do not have the time to vary-one-thing-at-a-time, a good approach might be to simultaneously vary multiple variables at the same time and then apply the *win-stay, lose-shift* strategy (e.g., Nowak & Sigmund, 1993), in which a strategy is pursued in case of success and only modified in case of failure.

Common trait variance of various strategies such as VOTAT or *win-stay, lose-shift* strategy could be modeled as latent second order factor representing metastrategic knowledge. If the trait specificity of this factor is high, participants who have high scores on that metastrategic knowledge factor are good in matching strategies to requirements of tasks. Measuring *use of strategies* in several tasks and modeling a second order factor may cover the CPS competency to search for information in an unknown environment more broadly than solely assessing single strategies within CPS tests.

Finally, in this thesis, *use of strategies* only captures the competency of using strategies for *knowledge acquisition*, not for *knowledge application*. According to Schoppek (2004), the

difference between these strategies is that *knowledge acquisition* strategies such as VOTAT focus on the effect of input variables (i.e., which effect does a certain input variable have), whereas *knowledge application* strategies focus on output variables (i.e., how is a certain output variable affected). He suggests measuring a common *knowledge application* strategy, in which the problem solver first has to predict the next states of the environment in case of no interventions, then the problem solver has to calculate differences between predicted and desired states, to select input variables for variation, to calculate values for input variables, and finally to apply the strategy. Measuring such knowledge application strategies using process data would foster understanding of problem solving processes in *knowledge application* and is therefore highly recommended for future research.

(5) *Longitudinal studies*: To evaluate development of CPS competency across age, longitudinal studies combined with training studies will lead to a more consistent picture allowing the interpretation of intra-individual differences across time. In Paper 2 and Paper 3, performance differences were measured across problem solvers of different ages in a cross-sectional design and it was mentioned that training efficient domain-general problem solving strategies may enhance participants understanding of how to generate and apply knowledge. Research already showed that participants who were trained in applying VOTAT in a science problem are able to transfer their knowledge and apply VOTAT also in other science tasks (e.g., Chen & Klahr, 1999; Klahr, Triona, & Williams, 2007). As outlined in Paper 4, MicroDYN tasks allow the illustration of the usefulness of domain-general strategies (e.g., VOTAT), because they can be embedded in various contexts. This emphasizes the utility of MicroDYN as a training tool for domain-general strategies such as VOTAT or strategies to identify interaction effects.

By analysing intra-individual developments of applying domain-general strategies across time and by comparing developmental differences across participants (i.e., analysing inter-individual differences in intra-individual change), prerequisites for students'

understanding of the usefulness of a certain strategy can be identified. Going one step further, inter-individual differences in intra-individual change can also be studied simultaneously for different methods. By applying a longitudinal MTMM model developed by Geiser (2009), it could be tested if working with interactive MicroDYN tasks is more useful to train domain-general strategies than verbal explanations or video instructions.

Finally, the role of emotion and motivation in CPS (e.g., Barth & Funke, 2010), cognitive modelling (Anderson, 2007), adaptive testing (Eggen, 2008), or collaborative problem solving (Greiff, in press) are not considered in this thesis, but nevertheless important from a theoretical and practical point of view, to name only some of the most interesting venues for future research on MicroDYN and MicroFIN that have not been mentioned yet.

6.4 Conclusion

This thesis dealt with the assessment of CPS, that is, the domain-general competency to generate and apply new knowledge while interacting with a previously unknown environment. Overall, data analyses confirmed that MicroDYN is suitable for measuring CPS not only in samples of university students, but also across participants with a broad range of cognitive performances including high school students and blue-collar workers. Further, CPS showed incremental validity beyond traditional measures of general mental ability. From a scientific perspective, these results emphasize the potential of MicroDYN as a measurement device of CPS and of the construct itself in explaining cognitive performance.

From a practical point of view, it seems to be undisputed that CPS competency is frequently needed in daily life. For instance, the economists Autor, Levy, and Murnane (2003) showed that due to the increasing utilization of technology in the last few decades, working people have to carry out more and more non-routine problem solving tasks instead of applying a well-defined set of cognitive and manual routine activities. In other words, in the 21st century, it is not sufficient to only apply factual knowledge to solve routine tasks. It is the

competency of generating and applying new knowledge that enables people to meet demands at the work place.

This is also acknowledged by the Organisation for Economic Co-operation and Development (OECD) which considers CPS as a valuable aspect of school achievement. The OECD integrated CPS in the Programme for International Student Assessment (PISA) in the 2012 cycle (OECD, 2010), emphasizing the interest in students' domain-general CPS competency. In educational psychology, it is confirmed that if "we only rely on teaching people the knowledge base or expertise [...], they [students] will not learn to be good general problem solvers" (Tuckman & Monetti, 2010, p.328). Further, Tuckman and Monetti (2010) suggest teaching domain-general strategies to increase general problem solving performance. However, first it has to be understood which conceptions and misconceptions exist when students use strategies to generate and apply knowledge. Thus, an adequate assessment of students' existing CPS competency is a prerequisite for teaching students to become better problem solvers. In this respect, a first step has been taken with the development of MicroDYN. Bearing in mind that people constantly interact with dynamically changing environments in the 21st century, solely using static tasks to measure cognitive performance does not meet the requirements of modern assessment. The world moves on and assessment of relevant competencies such as CPS has to keep pace.

References

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitschke, A., & Funke, J. (2012). Die Bedeutung übergreifender kognitiver Determinanten für die Bewältigung beruflicher Anforderungen. Untersuchung am Beispiel dynamischen und technischen Problemlösens. *Zeitschrift für Erziehungswissenschaft, 15*(2), 363-391.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

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

Barth, C. M. & Funke, J. (2010). Negative affective environments improve complex problem solving performance. *Cognition and Emotion, 24,* 1259-1268.

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

Byrne, B. M. & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*(2), 287-321.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 54,* 297-312.

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.

Cherney, I. D., & London, K. (2006). Gender-linked differences in the toys, television shows, computer games, and outdoor activities. *Sex Roles, 54,* 717-726.

Chi, M., & VanLehn, K. (2007) The impact of explicit strategy instruction on problem-solving behaviros across intelligent tutoring systems. In D. McNamara & G. Trafton (Eds.) *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* pp. 167-172. New York, NY: Erlbaum.

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.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Strucutural equation models for multitrait-multimethod data. In M. Eid & E Diener (Ed.), *Handbook of multimethod measurement in psychology* (pp. 283-299). Washington, DC: APA.

Eggen, T. J. H. M. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (p. 215-234). Göttingen: Hogrefe.

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

Geiser, C. (2009). *Multitrait-Multimethod-Multioccasion Modeling.* Akademische Verlagsgemeinschaft München: München.

Greiff, S. (in press). From Interactive to Collaborative Problem Solving: Current issues in the Programme for International Student Assessment. *Review of Psychology.*

Greiff, S. (2012). *Individualdiagnostik der Problemlösefähigkeit*. [Diagnostics of problem solving ability on an individual level]. Münster: Waxmann.

Greiff, S., Wüstenberg, S., Exner, K., Hofmann, F., Zweck, B., Mehlau, A., & Funke, J. (2010). *MicroFIN: The concept of finite automatons and its application within the assessment of problem solving competencies*. Heidelberg: Department of Psychology.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement*, *36*(3), 189-213.

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

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

Kuhn, D. (2000). Metacognitive development. *Current directions in Psychological Science, 9*(5), 178-181.

Kuhn, D. (2011). What is scientific thinking and how does it develop? *The Wiley-Blackwell handbook of childhood cognitive development (2nd ed.).* (pp. 497-523): Wiley-Blackwell.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*(4), 495-523.

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

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature, 364,* 56–58.

OECD (2010). *PISA 2012 Problem Solving Framework.* Paris: OECD Publishing.

Richter, T., Naumann, J., & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R) [A revised version of the INCOBI-R inventory for assessing computer knowledge]. *Zeitschrift für Pädagogische Psychologie, 24*(1), 23–37.

Schoppek, W. (2004). Teaching structural knowledge in the control of dynamic systems: Direction of causality makes a difference. In K. D. Forbus, D. Gentner, & T. Regier

(Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1219-1224). Mahwah, NJ: Lawrence Erlbaum Associates.

Schweizer, F., Wüstenberg, S., & Greiff, S. (in press). Do Complex Problem Solving dimensions measure something beyond working memory capacity? *Learning and Individual Differences.*

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality, 13*, 389-408.

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.

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology, 82*(2), 306-314.

Tuckman, B. W. & Monetti, D. M. (2010). *Educational Psychology.* Belmont: Wadsworth Cengage Learning.

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.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 - Revision - (CFT 20-R)* [Culture Fair Intelligence Test 20-R – Scale 2]. Göttingen: Hogrefe.

Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55-72). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Erklärungen

Erklärung gemäß § 8 Abs. 1 Buchst. b) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe.

Erklärung gemäß § 8 Abs. 1 Buchst. c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.

Eigenständige Beiträge in Paper 1 und Paper 4 betreffen anteilige Planung, Koordination und Durchführung der Erhebungen, Aufgabenprogrammierung sowie Federführung bei Auswertung und Verfassen beider Artikel.

Eigenständige Beiträge in Paper 2 und Paper 3 betreffen anteilige Planung, Koordination und Durchführung der Erhebungen, Aufgabenprogrammierung sowie substantielle Mitwirkung bei Auswertung und Verfassen beider Artikel.

Heidelberg, im Januar 2013 (Sascha Wüstenberg)

Curriculum Vitae

Persönliche Informationen

Name	Sascha Wüstenberg
Adresse	Schlossstraße 4a
	76646 Bruchsal
E-Mail	Sascha.Wuestenberg@t-online.de
Geburtsdatum	03.10.1982
Geburtsort	Waiblingen

Ausbildung

08.1993 – 06.2002	Limes-Gymnasium Welzheim, Abitur
09.2002 – 06.2003	Zivildienst beim Verein für Behinderte Schorndorf
10.2003 – 03.2004	Studium der Psychologie, Soziologie und vergleichende Religionswissenschaft (Magister), Universität Heidelberg
04.2004 – 10.2009	Studium der Diplom-Psychologie mit Wahlfach Betriebswirtschaftslehre, Universität Heidelberg

Wissenschaftliche Tätigkeiten

seit 11.2009	Wissenschaftlicher Mitarbeiter und Doktorand an der Universität Heidelberg (50%), Fachbereich Allgemeine Psychologie
08.2010 – 09.2011	Wissenschaftlicher Mitarbeiter an der Pädagogischen Hochschule Heidelberg (25%), Fachbereich Geographie