Learning and Individual Differences 29 (2014) 18–29

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

# Cross-national gender differences in complex problem solving and their determinants

Sascha Wüstenberg [a,*], Samuel Greiff [a], Gyöngyvér Molnár [b], Joachim Funke [c]

[a] University of Luxembourg, Luxembourg
[b] University of Szeged, Hungary
[c] University of Heidelberg, Germany

ARTICLE INFO

*Article history:*
Received 20 July 2012
Received in revised form 18 July 2013
Accepted 11 October 2013

*Keywords:*
Complex problem solving
MicroDYN
Measurement invariance
Gender differences
Nationality differences

ABSTRACT

The present study examined cross-national gender differences in domain-general complex problem solving (CPS) 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 strategy, which lead to considerably worse knowledge acquisition. Results imply that analyzing process data such as use of strategies is highly advisable to identify determinants of overall performance differences in CPS across groups of interest.



## 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 led to changes of the educational system and revisions of educational standards in Germany (Wernstedt & John-Ohnesorg, 2009), because German students underperformed in comparison to neighboring countries. Especially performance differences in domain-specific areas such as mathematical ability play an important role not only 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. For instance, domain-general problem solving competency was assessed in the 2012 cycle of PISA, which was conducted by the Organisation for Economic Co-operation and Development (OECD) with results scheduled for publication in 2014. 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 competency for students that should be developed in addition to domain-specific knowledge in school subjects. Within domain-general problem solving, (non-interactive) analytical problem solving and (interactive) complex problem solving (CPS) can be distinguished as subordinate constructs (Fischer, Greiff, & Funke, 2012; OECD, 2010). Whereas analytical problem solving is usually measured with static paper–pencil tasks, complex problem solving (CPS)[1] includes tasks enabling interactions between user and task situation (Wirth & Klieme, 2003). Recently, the OECD emphasized the importance of the domain-general competency to interactively deal with novel problems:

> *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]

* Corresponding author at: University of Luxembourg, 6, rue Richard Coudenhove Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg. Tel.: +352 466644 9704.
*E-mail address:* Sascha.Wuestenberg@uni.lu (S. Wüstenberg).

[1] The OECD used the term interactive problem solving (OECD, 2010) instead of CPS, referring to the interactive nature of the task. In the present paper, we use the term complex problem solving (CPS), which emphasizes the aspect of the underlying system's complexity. Both terms are used synonymously, but CPS is most established in research (Dörner, 1986, 1990; Funke, 2001, 2010).

1041-6080/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.lindif.2013.10.006

CPS tasks as described in this quotation usually contain many highly interrelated elements and system states of the tasks 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., whether a problem is solved or not), but also 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 problem solving are usually measured by two dimensions: 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 in domain-general CPS research (Bühner et al., 2008; Greiff, Wüstenberg, & Funke, 2012; Wüstenberg, Greiff, & Funke, 2012). Furthermore, CPS predicts supervisor ratings of participants' overall job performance (Danner, Hagemann, Schankin, Hager, & Funke, 2011) and school grade point average (Greiff & Fischer, 2013; Wüstenberg et al., 2012) even beyond reasoning. However, to our knowledge, no studies have yet been conducted analyzing individual differences in students' CPS performance and their determinants with regard to gender and nationality.

As a prerequisite for analyses and interpretations of overall differences in CPS performance and their determinants, it has to be ensured that there are no systematic demographic subgroup biases. For instance, in educational research on students' abilities as well as in intelligence testing, plenty of research has been conducted to ensure that measurement devices allow an unbiased measurement of the construct of interest across subgroups (e.g., Bowden, Saklofske, & Weiss, 2011; Chen, 2012; Gardner & Qualter, 2011). Especially performance differences with regard to gender and nationality often raise interest and concerns (cf. Else-Quest et al., 2010), leading to extensive discussions about determinants of performance, and – as outlined above in case of differing performance across nationality – to changes of whole educational systems (Wernstedt & John-Ohnesorg, 2009). Thus, it is vitally important to understand whether between-group differences in cognitive performance with regard to gender and nationality reflect true differences in the construct of interest, or different psychometric properties of the underlying measurement scale (Brown, 2006). But even if performance differences are valid, educationalists are not only interested in knowing *that* performance differences exists, but they also want to know *why* they exist in order to be able to foster the underlying competency by applying appropriate interventions. With regard to CPS, research on accurate measurement of performance differences across groups as well as determinants of performance is still in its infancy.

As will be further outlined, it is yet unclear whether CPS can be measured with equal validity across (1) gender or (2) nationality and analyses on individual differences in CPS performance are scarce. In fact, joint analyses of differences including (3) both gender and nationalities are non-existent. Particularly the latter is of high 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). There are also only few studies, which (4) investigate determinants of performance differences in CPS by analyzing process data gathered while participants interact with the task environment.

To this end, based on a sample of Hungarian and German high school students, (1) we will evaluate whether CPS can be measured with equal validity across gender and investigate gender differences in mean CPS performance. (2) We will analogously evaluate whether CPS can be measured with equal validity across Germans and Hungarians and investigate differences in mean CPS performance. (3) Further, we conduct combined analyses to study interaction effects of gender and nationality. Therefore, our sample is separated in four groups containing German males, German females, Hungarian males, and Hungarian females to evaluate whether CPS can be measured with equal validity across gender and nationality and to investigate mean differences across four groups. (4) Finally, we investigate determinants of mean performance differences across groups in knowledge acquisition by analyzing process data including behavioral patterns of participants gathered during their exploration of the tasks.

### 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; Sass, 2011), 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; and *Tailorshop*, Süß, 1996), no studies have been conducted analyzing measurement invariance with regard to gender. Only recently, it was shown that CPS can be measured invariant across Hungarian high school students in different grades (Greiff et al., 2013).

With regard to gender differences in CPS, previous findings are contrary to results on reasoning ability, in which reported gender differences slightly favor 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; Wittmann & Hattrup, 2004; Wittmann & Süß, 1999). However, the study of Jensen and Brehmer (2003) was based on a very small sample with limited generalizablity ($N = 15$; four males). As Wittmann and Hattrup (2004) integrated findings of Wittmann and Süß (1999) and two additional studies, we therefore only describe results of Wittmann and Hattrup (2004) in more detail.

Specifically, Wittmann and Hattrup (2004) pooled data of three independent studies using the CPS scenario *Tailorshop* (cf. Süß, 1996), 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 (Cohens' $d = .70$) 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 a lower amount of 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 proven 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 relevant for the problem situation.

(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 favoring 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, differences in CPS performance between males and females in the study of Wittmann and Hattrup (2004) might be partly 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 whether 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 MicroDYN, influence of prior knowledge on performance is minimized and semantic embedment of tasks is varied to ensure sufficient motivation of both males and females (cf. Section 2.2). This enhances the possibility that measurement invariance holds across groups, because subgroups cannot outperform each other due to different amounts of prior knowledge (cf. different knowledge about business contexts in *Tailorshop*) and, thus, task embedment should not advantage males or females.

As effect sizes of differences between males and females in Wittmann and Hattrup (2004) are too large to be solely explained by prior knowledge and motivation, and Wittmann and Hattrup (2004) showed that differences partly occurred by superior use of strategies in favor of males, we expect that males also perform better than females in MicroDYN. However, differences should be smaller than in the study of Wittmann and Hattrup (2004), because performance in MicroDYN tasks is less influenced by prior knowledge and motivation. When stating this, we acknowledge that our hypothesis concerning mean differences with regard to MicroDYN is rather explorative, because we can only rely on results of Wittmann and Hattrup (2004). They used *Tailorshop* as an assessment instrument of CPS, differing substantially from MicroDYN, for instance, by the role prior knowledge plays in it.

**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 and Cognitive Ability Test CogAt, Lakin, 2012). Investigating measurement invariance is particularly necessary when tests are applied in different countries. For instance, the underlying meaning of test items including verbal material may change during translation processes (Chen, 2008). As an example, 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. Further, participants may react differently to tasks due to cultural reasons. Thus, measurement invariance across nationalities has to be tested in order to ensure valid inferences about mean differences. We assume that measurement invariance holds across nationality in this paper, because bilingual native Hungarian interpreters translated the German version of the MicroDYN test into Hungarian, which should minimize translation errors. Further, we do not expect different behavioral patterns due to cultural reasons, because both German and Hungarian students are socialized in Western industrialized societies and are accustomed to classroom testing.

Equivalently to analyses on measurement invariance, studies dealing with cross-national mean differences in CPS performance are rather scarce. In a 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 Philippines, 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 of Hungarians and Germans, however, has only been conducted using paper–pencil tasks in the PISA 2003 assessment of problem solving (OECD, 2004). In this large-scale assessment, performance of Hungarian and German students did not differ significantly (OECD, 2004, p. 42). Whilst acknowledging that PISA is not a research venue and that the competency of actively generating information and using feedback required in CPS is not measured in paper–pencil tasks of problem solving (Buchner, 1995; Wüstenberg et al., 2012), we consider the PISA 2003 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 to CPS performance, we also analyze simultaneous interaction effects of both gender and nationality to establish a more detailed picture of CPS competencies and their determinants. Studies investigating such cross-national patterns of gender differences in problem solving performance have only been conducted using paper–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 Math 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 Math? 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 yields important information on how national characteristics (e.g., "status and welfare of women" and "differences within education systems", p. 125) are related to performance in specific domains. If gender differences in CPS 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, these analyses are 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 whether these results based on paper–pencil tasks can be readily applied to dynamic and interactive measures of CPS as these measures differ markedly from static paper–pencil test of problem solving that were used in PISA 2003 (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.

**Hypothesis 3b.** 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.

### 1.4. The impact of strategic behavior

Finally, we want to analyze how differences in mean performance across groups in knowledge acquisition are related to strategic behavior of participants applied during their interactions with the CPS tasks. To our best knowledge, there is only one study that investigated process data of cross-national differences in CPS performance from a strategic point of view (Strohschneider & Güss, 1999). Strohschneider and Güss (1999) reported that German problem solvers applied more control-oriented strategies than Indian problem solvers in the CPS task MORO. They also mentioned that Indian participants ignored more key aspects of the scenario, made more decisions without having the necessary information available, and forgot to control the effects of their treatments more often. Although results point towards considerable differences between nationalities in use of strategies, all variables used by Strohschneider and Güss (1999) refer to strategic behavior during knowledge application. However, in this paper, we want to analyze how participants' strategic behavior applied while exploring the CPS task predicts performance in knowledge acquisition.

An important indicator of performance in knowledge acquisition in CPS tasks is use of efficient strategies while exploring the task (Vollmeyer, Burns, & Holyoak, 1996). 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). Wüstenberg et al. (2012) could show empirically that applying VOTAT during the exploration phase in the CPS task MicroDYN, which was also applied in this study, is highly correlated with performance in knowledge acquisition in a highly selective sample of German university students. However, it has not been tested yet whether performance differences across groups are related to differences in use of VOTAT in a cross-national sample covering a broad range of cognitive ability.

**Hypothesis 4.** We expect that use of VOTAT explain performance differences across groups in knowledge acquisition.

From a theoretical point of view, it is reasonable that an efficient use of strategy such as VOTAT also influences performance in knowledge application via its effect on knowledge acquisition. That is, the better strategies are employed, the more knowledge is gathered, which in turn enables better performance in knowledge application. However, we are not interested in analyzing this indirect mediator effect and therefore did not test it.

## 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.[2] 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 samples, there were nearly as much females in each grade as males and gender distribution neither differed across grade levels 8 to 11 ($\chi^2 = 2.00$, df = 3, $p > .05$), nor across countries ($\chi^2 = 1.83$, 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), which was translated from German to

[2] 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., 2013).

Hungarian by bilingual native Hungarian interpreters for the test administration to the Hungarian speaking part of our sample. The MicroDYN approach uses several independent problems that rely on the formal framework of linear structural equations (cf. Appendix for equations) and can therefore be distinguished from semantically rich ad-hoc constructed simulations (for an overview see Funke, 2010). In MicroDYN, an entire set of independent tasks is employed with time-on-task being approximately 5 min for each task. Within a MicroDYN task (e.g., the task "perfume" shown in Fig. 1), input variables (e.g., fictitious ingredients labeled Norilan, Miral, and Carumin; upper left side of Fig. 1) influence output variables (e.g., flavors labeled Fresh, Fruity, and Flowery; upper right side of Fig. 1). Participants can actively manipulate input variables, whereas they can only observe changes in output variables. The procedure within a task is divided into (1) an exploration phase and (2) a control phase.

In the (1) exploration phase, relations between input and output variables are not visible to participants and they have to identify them by actively manipulating sliders of the input variables (time frame: 3.5 min). 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 Fig. 1). For instance, participants may draw an arrow between Norilan and Fresh. The CPS dimension knowledge acquisition is assessed by evaluating the correctness of the model drawn during the exploration phase.

Although the sample 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 demands on participants when exploring and controlling the task (Funke, 2001; Wüstenberg et al., 2012).

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 min). Targets are presented to participants by red areas and by numbers in brackets (upper right part of Fig. 1; target values are displayed only in the control phase). For instance, participants have to increase the value of Fresh by setting the slider of Norilan or Miral on "++". The CPS dimension knowledge application is assessed by evaluating whether target values are reached (for a detailed description of the MicroDYN approach see Greiff et al., 2012).

Test administration of MicroDYN started with a detailed instruction including a trial task, in which participants learned what they were expected to do in the exploration and control phase and how to deal with the software interface. Afterwards, participants worked on the MicroDYN tasks. Each task was embedded in different contexts to enhance motivation of students (e.g., training a handball team, mixing chemical elements, producing a perfume, and handling a moped). To avoid uncontrolled influences of prior knowledge, input or output variables were labeled either without deep semantic meaning (e.g., training A, B, and C for different training methods) 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).

In summary, MicroDYN minimizes uncontrolled influences of domain-specific prior knowledge that might have affected strategic behavior of participants while dealing with the tasks such as in *Tailorshop*.

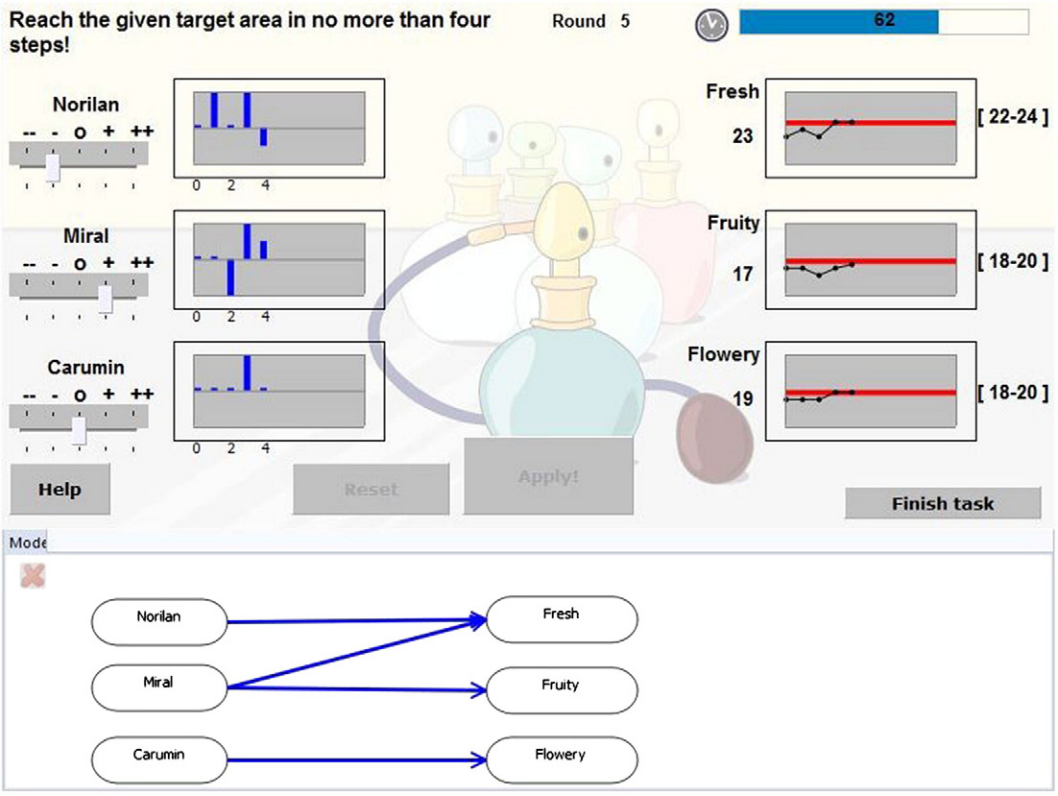


**Fig. 1.** Screenshot of the MicroDYN-task "perfume" during control phase. The sliders of the input variables range from "− −" (value = −2) to "++" (value = +2). Current values and target values are displayed graphically and numerically.

Further, MicroDYN allows the measurement of two core aspects of CPS, knowledge acquisition and knowledge application, by including multiple complex systems that constitute a recent measurement advancement in the field of CPS (Greiff et al., 2012). Empirical research has already shown that CPS as measured by MicroDYN is distinct from reasoning (Wüstenberg et al., 2012) and working memory (Schweizer, Wüstenberg, & Greiff, 2013), and incrementally predicts relevant outcomes such as school grades (Wüstenberg et al, 2012). Further, recent results indicate that MicroDYN allows an invariant measure of CPS across high school students in different grades (Greiff et al., 2013). Taken together, MicroDYN seems suitable for the measurement of CPS, although important research questions such as invariance with regard to gender and nationality are yet to be tested. However, the MicroDYN approach suffers from some limitations that we further elaborate in the discussion (cf. Section 4.4).

### 2.3. Dependent variables

Knowledge acquisition, knowledge application, as well as use of VOTAT was scored dichotomously (see Greiff et al., 2012; Kröner et al., 2005). For knowledge acquisition, 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, credit was given if target values of all variables were reached and no credit, if at least one target value was not reached. For use of VOTAT, credit was given if participants applied VOTAT for all input variables and no credit, if VOTAT was used inconsistently or not at all.

### 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 min. 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. 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).

With regard to measurement invariance, the first step is to identify a baseline model that fits within the overall sample and in each subgroup (Byrne & Stewart, 2006). 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 (Hypotheses 1a and 1b; male vs. female), nationality (Hypotheses 2a and 2b; Germans vs. Hungarians), and gender and nationality (Hypotheses 3a and 3b; 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 (i.e., manifest outcomes were scored right or wrong). Consequently, constraints on model parameters differed in comparison to invariance tests based on continuous outcomes. Thus, in all analyses, we started testing configural invariance by estimating the parameters of the baseline model once again in a multi-group 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 $\chi^2$-difference test (cf. Byrne & Stewart, 2006; Muthén & Muthén, 2010). If the $\chi^2$-difference test is non-significant, measurement invariance exists. Beyond testing invariance on an overall level, additionally we applied Lagrange-Multiplier tests (LM) to check whether a certain group had advantages in specific tasks. 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 specific task, for which the factor loading or the threshold was freed, 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 indicated a good global model fit, the model showed a good fit in the overall sample ($\chi^2 = 129.073$, $df = 40$, $p < .001$; CFI = .975, RMSEA = .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 ($\chi^2 = 229.144$, $df = 41$, $p < .001$; CFI = .962, RMSEA = .072) with knowledge acquisition and knowledge application combined under one factor as indicated by a significant $\chi^2$-difference test ($\chi^2 = 74.489$; $df = 1$; $p < .001$).

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 = .968 to .985, RMSEA = .022 to .054), and, thus, was used as baseline in

each of the following analyses (Hypotheses 1a to 3b). Communalities for knowledge acquisition ($h^2$ = .44–.82) and knowledge application ($h^2$ = .26–.79) were mostly above the recommended level of .40 (Hair, Anderson, Tatham, & Black, 1998). However, due to the comparable low number of only six administered tasks, internal consistencies were smaller (knowledge acquisition $\alpha$ = .74; knowledge application $\alpha$ = .62) than in other studies using MicroDYN tests based on larger task samples (e.g., Greiff et al., 2012; Wüstenberg et al., 2012).

### 3.2. Hypotheses 1a and 1b: 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 two rows in Table 1 labeled gender). Thus, CPS is measurement invariant across gender, supporting Hypothesis 1a.

We applied LM-tests to ensure that embedment in a certain context used within a MicroDYN task did not unjustifiably favor 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) than females, supporting Hypothesis 1b.

### 3.3. Hypotheses 2a and 2b: Nationality

We assumed that CPS is measured invariant with regard to nationality. Results showed that measurement invariance held (see third and fourth row in Table 1 labeled nationality) 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 1a, 1b and 2a, 2b, 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. Hypotheses 3a and 3b: Nationality and gender

In order to allow for 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 (Hypotheses 3a and 3b). 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 fifth and sixth rows in Table 1 labeled nationality and gender) and LM tests did not yield any significant result, supporting measurement invariance for the four subgroups as expected within 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).

Results showed that German males performed significantly better in knowledge acquisition and knowledge application than both Hungarian and German females, but differed non-significantly from Hungarian males. 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. Taken together, German and Hungarian males performed best, followed by German females who differed significantly from both groups in knowledge application, but non-significantly from at least Hungarian males in knowledge acquisition. Hungarian females underperformed considerably in contrast to all other groups in both dimensions.

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 were 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 content knowledge is not necessary to solve the MicroDYN tasks; cf. Section 2.2) sufficiently explained these performance differences. That is, latent mean differences are likely to display real differences in underlying CPS performance, but it remains unclear how these differences can be explained by actual behavior when working on MicroDYN. This question is tackled in Hypothesis 4, in which we investigated use of VOTAT as determinant of performance differences in knowledge acquisition across groups.

### 3.5. Hypothesis 4: The impact of strategic behavior

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. Descriptive analyses based on the overall sample showed that internal consistency of applying VOTAT across tasks was good ($\alpha$ = .89). Mean use of VOTAT was highest for German males ($M_{votat}$ = .76, SD = .32), followed by Hungarian males ($M_{votat}$ = .62, SD = .37) and German females ($M_{votat}$ = .62, SD = .39), but was considerably lower for Hungarian females ($M_{votat}$ = .35, SD = .39). One-way

**Table 1**
Goodness of fit indices for measurement invariance of CPS.

| Group | Invariance model | $\chi^2$ | df | p | CFI | TLI | RMSEA | Free par. | Compare with | $\Delta\chi^{2a}$ | $\Delta df^a$ | 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 and 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.
$\chi^2$ and df are estimated by WLSMV.
$\chi^2$-differences cannot be compared by subtracting $\chi^2$ and df if WLSMV-estimators are used.
[a] $\Delta\chi^2$ and $\Delta$df were estimated by a $\chi^2$-difference test procedure in MPlus (see Muthén & Muthén, 2010).

analyses of variances showed significant differences across groups ($F(3, 881) = 55.663$, $p < .001$). We conducted three linear contrasts revealing that females significantly applied VOTAT less often than males ($t(881) = 8.460$, $p < .001$), Hungarians used it significantly less often than Germans ($t(881) = 8.580$, $p < .001$), and so did Hungarian females in comparison to the three other groups ($t(881) = 12.180$, $p < .001$). Thus, manifest analyses indicated that differences in knowledge acquisition performance were reflected in differences in use of VOTAT.

This is also supported by a latent mediator analyses, in which gender served as bivariate predictor variable, use of VOTAT as a latent factor mediator (incorporating one indicator per task for use of VOTAT), and knowledge acquisition as latent dependent factor. Results of the adequate fitting mediator model ($\chi^2 = 318.380$, df = 40, $p < .001$; CFI = .980, RMSEA = .088) showed that there was no direct effect from gender to knowledge acquisition (unstandardized path coefficient: b = .04, *SE* = .04, p = .27).[3] However, gender was significantly related to the mediator use of VOTAT (b = .47, *SE* = .08, p < .001) and use of VOTAT was in turn strongly related to knowledge acquisition as dependent variable (b = .97, *SE* = .06, p < .001). Further, the total indirect mediator effect from gender to knowledge acquisition via use of VOTAT was significant (b = .45, *SE* = .06, $p < .001$). Results with nationality as mediator were comparable (not reported), also showing a significant total indirect mediator effect from nationality to knowledge acquisition via use of VOTAT.

In summary, results clearly showed that performance differences across groups in knowledge acquisition strongly depended on use of VOTAT, which explains why females outperformed males, Hungarians outperformed Germans, and Hungarian females underperformed compared to all other groups in knowledge acquisition.

## 4. Discussion

The general aim of this study was to examine whether 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. Measurement invariance held in all analyses (cf. Section 4.1). Further, results revealed that males outperformed females, Germans outperformed Hungarians, and, interestingly, Hungarian females performed worse than all other groups (cf. Section 4.2). Finally, analyses on logfile data showed that performance differences in knowledge acquisition strongly depended on use of VOTAT. We conclude that it is essential to investigate use of efficient strategies to obtain a detailed picture of determinants of performance in CPS (cf. Section 4.3).

[3] We reported only unstandardized path coefficients, because MPlus does not provide SE and p-values for standardized estimates of β-coefficients when WLSMV estimator for categorical variables is used in a mediator model with covariates as we did in our analyses.

### 4.1. Measurement invariance

As expected, results provided support that CPS was 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: Firstly, latent mean differences can be meaningfully interpreted (Byrne & Stewart, 2006). Secondly, 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 minimized, do not favor 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). Thirdly, results on measurement invariance across nationality showed 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 indicated real differences in underlying CPS competency and were unlikely to be a methodological artifact.

### 4.2. Latent mean comparisons

Latent mean comparisons showed that males outperformed females as expected (Hypothesis 1b), but contrarily to our hypothesis there was a significant difference between Germans and Hungarians in favor of Germans (Hypothesis 2b) in both CPS dimensions. Comprehensive comparisons of all four groups revealed an interaction effect in cross-national performance differences showing that Hungarian females performed significantly worse than all other groups.

### 4.3. Determinants of performance: Use of VOTAT

Further analyses on determinants of performance revealed that performance differences in Hypotheses 1a through 3b in knowledge acquisition 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. As mentioned in the introduction, applying strategies such as VOTAT is a prerequisite for gaining knowledge in CPS tasks as it is assessed in knowledge acquisition (Kröner et al., 2005), explaining overall performance differences between groups in this dimension by exploration behavior. As shown by Wüstenberg et al. (2012), use of VOTAT in the exploration phase also affects performance in knowledge application. That is, use of efficient strategies yields in more generated knowledge, which in turn influences performance in knowledge application. This also applies if the correct model is given in the control phase (Wüstenberg et al., 2012). Nevertheless, we acknowledge that we did not test this

**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.

effect in the present paper and we also did not investigate specific strategies used in the phase of knowledge application (cf. Section 4.4, Limitations).

Our results on gender differences partly coincide with findings of Wittmann and Hattrup (2004). In their study, males outperformed females because they altered variables 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 may indicate that males revert to more efficient strategies when building up knowledge while dealing with an unknown problem, 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 (*ds* = .20–.25) and Hungarians (*ds* = .38–.43) in both CPS facets, showing a clear interaction effect of gender and nationality. However, the large differences in performance especially between Hungarian females and other groups were surprising and not a priori expected.

This leads to the question why (Hungarian) females applied VOTAT less often than other groups. One possible explanation among others is a missing understanding of the concept of additive effects from input variables on output variables. That is, "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 and 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, implying that this explanation remains speculative until it is empirically supported. However, it could be an interesting venue for future research to analyze potential determinants of appropriate use of strategies such as understanding additive effects.

The fact that some groups used less VOTAT might also be explained by the specific way schooling is set up in a particular educational system. In school, understanding and applying the principle of VOTAT is commonly taught in science, because VOTAT allows a logical disconfirmation of alternative hypotheses, which is central to most experimental designs (Kuhn et al., 2000; Tschirgi, 1980; 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 that can be used for teaching domain-general strategies such as VOTAT may be 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, but this hypothesis is yet to be tested. In summary, this study showed for the first time that large differences across groups in overall performance on knowledge acquisition were directly related to differences in use of VOTAT. However, the reasons for the differences in application of VOTAT remained unclear and we provided only preliminary suggestions for potential causes, implying that specific explanations of these effects have yet to be revealed.

### 4.4. Limitations

There are clearly some limitations in this study that may guide future research. In this section, we want to focus on three of these limitations: (1) Applying MicroDYN tasks to measure CPS narrowed the construct to some extent (MicroDYN represents only one possibility to measure CPS); (2) only little information was available from previous research for deriving hypotheses with regard to performance differences across gender and nationality, which yielded in exploratory hypotheses in this study; and (3) we did not measure use of strategies within knowledge application.

Ad (1): MicroDYN allows the measurement of some core features of CPS, that is, the assessment of knowledge acquisition and knowledge application in interactive and dynamically changing environments. However, compared to other CPS tasks that try to simulate highly complex real world situations such as *Tailorshop* (cf. Wittmann & Süß, 1999), the complete system structure can be identified in tasks based on linear structural equation systems such as MicroDYN. In *Tailorshop*, a large amount of highly interconnected intervening variables are included that are partly not visible to the problem solver, which implies that participants have to deal with a scenario that they cannot fully describe in terms of a correct causal model. This component of dealing with incomplete information and uncertainty, which could also be considered as a part of CPS competency, is not integrated in the assessment within MicroDYN.

Further, MicroDYN rests upon multiple linear structural equation systems comparable to other established CPS scenarios (i.e., CogSIM, Kluge, 2008; *MultiFlux*, Kröner et al., 2005) that only allow including quantitative relations between input and output variables. 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 or are of a general nature. Finite state automata tasks rely on qualitative connections between variables and can represent a large variety of devices encountered in

everyday life (e.g., ticket vending machines and mobile phones; Funke, 2001; Funke & Frensch, 2007). 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 such as MicroDYN in order 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, which was based on use of the VOTAT strategy. This would tell us whether males use more efficient strategies in CPS tasks in general or whether this result is limited to use of specific strategies such as VOTAT.

Ad (2): As already mentioned in the Introduction, this paper is the first dealing with performance differences in CPS across nationality and we could only rely on findings of one paper dealing with mean gender differences in CPS (i.e., Wittmann & Hattrup, 2004). Thus, analyses of mean differences were of a somewhat explorative nature. Nevertheless, we did build upon findings of Wittmann and Hattrup (2004) by showing that differences in performance across gender and nationality were influenced by differences on use of strategies. In contrast to the study of Wittmann and Hattrup (2004) using the *Tailorshop*, however, performance differences in our study could be clearly related to use of strategies, because domain-specific knowledge does not strongly affect performance in MicroDYN.

Ad (3): With regard to use of strategies, we only measured the competency of choosing appropriate strategies in the exploration phase. Although use of appropriate strategies may indirectly affect knowledge application above and beyond its effect on knowledge acquisition (cf. Wüstenberg et al., 2012), we did not explicitly measure application strategies in the control phase. We showed that applying VOTAT in the exploration phase leads to better performance in knowledge acquisition explaining performance differences across groups in a mediator analysis. However, it remains unclear which further strategies during the application of knowledge may lead to performance differences in overall performance of knowledge application. According to Schoppek (2004), 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). Within MicroDYN, for instance, output variables (e.g., "Fresh", cf. Fig. 1) that are influenced by multiple input variables (e.g., "Norilan" and "Miral", cf. Fig. 1) should be considered first when participants are asked to achieve target goals, because the value of output variables with multiple effects (e.g., "Fresh") will also change when an input variable linked to them (e.g., "Miral") is manipulated to control another output variable (e.g., "Fruity", cf. Fig. 1). This kind of strategy is not only restricted to MicroDYN, but also ensures better performance in other CPS scenarios such as ColorSIM (Kluge, 2008) or *MultiFlux* (Kröner et al., 2005). In this respect, it would be worthwhile analyzing whether those participants who pay more attention to output variables that are affected by many input variables perform better in knowledge application than those participants not doing this. 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.

### 4.5. Outlook

In recent years, education seeks to foster competencies that are relevant in a number of domains besides domain-specific knowledge. These efforts proceed on the general assumption of transfer taking place in education (Perkins & Salomon, 1989) and that students are able to use what they learned in one domain also in other domains (OECD, 2010). As an example, applying useful strategies such as VOTAT is relevant in identifying causal relations between variables in various domains (e.g., biology, economics, physics, and psychology) and is, therefore, of high importance in educational contexts. For instance, in both physics and psychology, only the variable of interest is varied in experimental designs while all other variables that may also influence results are held constant. In this respect, MicroDYN offers great opportunities, because it can be used to teach domain-general strategies such as VOTAT. According to Adam (1989), the more specific the context is in which thinking skills are trained or knowledge is acquired, the lower the possibility of transferring 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, which may allow an easier transfer to other domains. Thus, future research may investigate usefulness of MicroDYN not only as an assessment instrument, but also as a training tool for teaching domain-general strategies such as VOTAT.

However, in order to enable transfer of knowledge, learners must understand when the application of what has been learned is useful (Bransford, Brown, & Cocking, 1999). With regard to use of strategies, this aspect of meta-cognition is called meta-strategic knowledge. That is, the competency 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 meta-strategic 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 therefore 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 meta-strategic knowledge of students. This study can be regarded as a first step as it clearly showed that identification of one important strategy applied by participants (i.e., VOTAT) explained their performance in knowledge acquisition. However, analyzing process data gathered during students' interactions with various CPS tasks would allow a deeper understanding of cognitive processes engaged.

### 4.6. Conclusion

In the present study, we showed that CPS assessed through MicroDYN is measured invariant across gender and across two nationalities and that these groups differ in their overall mean performance. We further investigated the determinants of mean differences across groups, showing that analyses should not only be limited to outcome variables, but also focus more on process variables such as use of strategies (e.g., VOTAT) stored in log files. In fact, large-scale studies describe and compare overall performance in a number of different competencies, so happened for CPS in the PISA 2012 cycle. Description of such differences is a first important step and provides a benchmark for cross-national comparisons. However, the even more important question for people involved in education is whether competencies such as CPS – in particular when they cannot be allocated within specific school subjects – can be fostered. That is, we need to gain a deeper understanding of the specific behaviors leading to mean group performance differences and the mechanisms that can be used to change these behaviors. We therefore suggest that further research should investigate reasons for differences in CPS also from an educational teaching perspective. Today's students shape the world of tomorrow and "society expects that the problem-solving lessons learned in school […] will transfer to students' adult lives for the betterment of the world" (Novick & Bassok, 2005, p.345). Analyzing log file data as well as the specific behavioral strategies stored in them will strongly contribute to this goal — an opportunity that should not be missed.

## 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 | $X_{t+1} = 1 \times X_t + 0 \times A_t + 2 \times B_t$<br>$Y_{t+1} = 1 \times Y_t + 0 \times A_t + 2 \times B_t$ | 2 × 2-system | Only effects of inputs |
| Task 2 | $X_{t+1} = 1 \times X_t + 2 \times A_t + 2 \times B_t + 0 \times C_t$<br>$Y_{t+1} = 1 \times Y_t + 0 \times A_t + 0 \times B_t + 2 \times C_t$ | 2 × 3-system | Only effects of inputs |
| Task 3 | $X_{t+1} = 1 \times X_t + 2 \times A_t + 0 \times B_t + 0 \times C_t$<br>$Y_{t+1} = 1 \times Y_t + 0 \times A_t + 2 \times B_t + 2 \times Ct$<br>$Z_{t+1} = 1 \times Z_t + 0 \times A_t + 0 \times B_t + 2 \times C_t$ | 3 × 3-system | Only effects of inputs |
| Task 4 | $X_{t+1} = 1 \times X_t + 2 \times A_t + 2 \times B_t + 0 \times C_t$<br>$Y_{t+1} = 1 \times Y_t + 0 \times A_t + 2 \times B_t + 0 \times C_t$<br>$Z_{t+1} = 1 \times Z_t + 0 \times A_t + 0 \times B_t + 2 \times C_t$ | 3 × 3-system | Only effects of inputs |
| Task 5 | $X_{t+1} = 1.33 \times X_t + 2 \times A_t + 0 \times B_t + 0 \times C_t$<br>$Y_{t+1} = 1 \times Y_t + 0 \times A_t + 0 \times B_t + 2 \times C_t$ | 2 × 3-system | Effects of inputs and outputs |
| Task 6 | $X_{t+1} = 1 \times X_t + 2 \times A_t + 0 \times B_t + 0 \times C_t$<br>$Y_{t+1} = 1 \times Y_t + 2 \times A_t + 0 \times B_t + 0 \times C_t$<br>$Z_{t+1} = 1.33 \times Z_t + 0 \times A_t + 0 \times B_t + 2 \times C_t$ | 3 × 3-system | Effects of inputs and 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.

## 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.

Bowden, S.C., Saklofske, D. H., & Weiss, L. G. (2011). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-IV in the United States and Canada. *Educational and Psychological Measurement, 71*(1), 186–199.

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: Guildford 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. (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. (2012). Measurement invariance of WISC-IV across normative and clinical samples. *Personality and Individual Differences, 52*(2), 161–166.

Chen, H., Keith, T. Z., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, 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.

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.

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.

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 in hazardous situations* (pp. 15–36). New York, NY: Oxford University Press.

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.

Gardner, K. J., & Qualter, P. (2011). Factor structure, measurement invariance and structural invariance of the MSCEIT v2.0. *Personality and Individual Differences, 51*(4), 492–496.

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

Greiff, S., & Fischer, A. (2013). Der Nutzen einer Komplexen Problemlösekompetenz: Theoretische Überlegungen und empirische Befunde [Usefulness of complex problem solving competency: Theoretical considerations and empirical results]. *Zeitschrift für Pädagogische Psychologie, 27*(1), 27–39.

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. (2013). Complex problem solving in educational settings — Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*, 364–379.

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.

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.

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 ed.)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* (pp. 321–349). Cambridge, NY: University Press.

OECD (2004). *Problem solving for tomorrow'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.
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. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning & Individual Differences*, *24*, 42–52.
Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., et al. (2012). The genetics lab. Acceptance and psychometric characteristics of a computer-based microworld to assess complex problem solving. *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.