Journal of Educational Psychology

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

Samuel Greiff, Sascha Wüstenberg, Gyöngyvér Molnár, Andreas Fischer, Joachim Funke, and Benő Csapó

Online First Publication, February 18, 2013. doi: 10.1037/a0031856

Journal of Educational Psychology

0022-0663/13/$12.00 DOI: 10.1037/a0031856

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

Samuel Greiff and Sascha Wüstenberg
University of Luxembourg

Gyöngyvér Molnár
University of Szeged

Andreas Fischer and Joachim Funke
University of Heidelberg

Benő Csapó
University of Szeged

Innovative assessments of cross-curricular competencies such as complex problem solving (CPS) have currently received considerable attention in large-scale educational studies. This study investigated 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. Hungarian high school students in Grades 5 to 11 ($N = 855$) 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.

Keywords: complex problem solving, general mental ability, intelligence, MicroDYN, education

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 of 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:

Samuel Greiff and Sascha Wüstenberg, EMACS Unit, University of Luxembourg, Luxembourg-Kirchberg, Luxembourg; Gyöngyvér Molnár, Institute of Education, University of Szeged, Szeged, Hungary; Andreas Fischer and Joachim Funke, Department of Psychology, University of Heidelberg, Heidelberg, Germany; Benő Csapó, Institute of Education, University of Szeged, Szeged, Hungary.

This research was funded by grants supported by the Fonds National de la Recherche Luxembourg (ATTRACT; ASSKI21) and by the German Federal Ministry of Education and Research (LSA004 and 01JG1062). We are grateful to the Technology Based Assessment group at DIPF (http://tba.dipf.de) for providing the authoring tool CBA Item Builder and technical support.

Correspondence concerning this article should be addressed to Samuel Greiff, EMACS Unit, University of Luxembourg, 6 rue Richard Coudenhove Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg. E-mail: samuel.greiff@uni.lu

[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 article, we use the term *complex problem solving* (CPS), which is the most established in research.

> 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 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 (1904) 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 labeled 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 both to predict and to understand complex learning processes in classrooms (Mayer & Wittrock, 2006). Similar to Sternberg and his conception of intelligence (Sternberg, 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 domain-specific 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 domain-specific 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. Not only is this delineation into two main cognitive processes logical and 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 means-ends analysis (Newell & Simon, 1972). In fact, Greeno and Simon (1988) stated that problem solvers tend to prefer a means-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 pencil-and-paper 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). 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 labeled 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 labeled *blue, green,* or *red chips* serve as input variables, whereas different kinds of playing cards labeled *Royal, Grande,* or *Nobilis* serve as output variables.

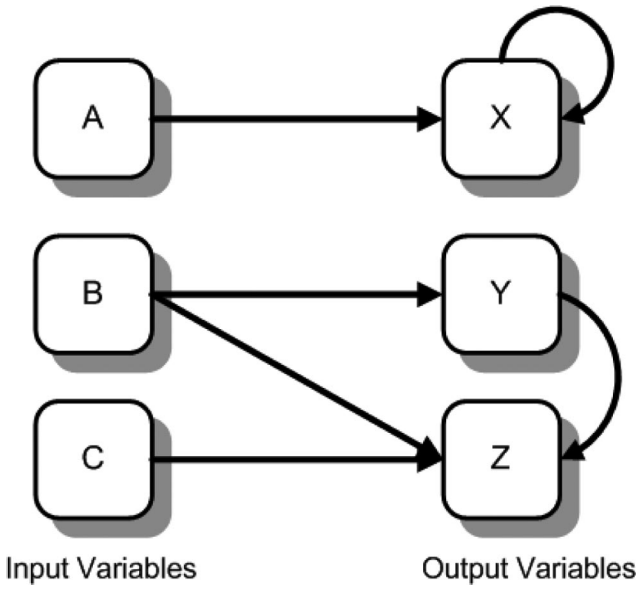

Figure 1. Structure of a typical MicroDYN task displaying three input (*A, B, C*) and three output (*X, Y, Z*) variables.

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 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 are 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 come 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

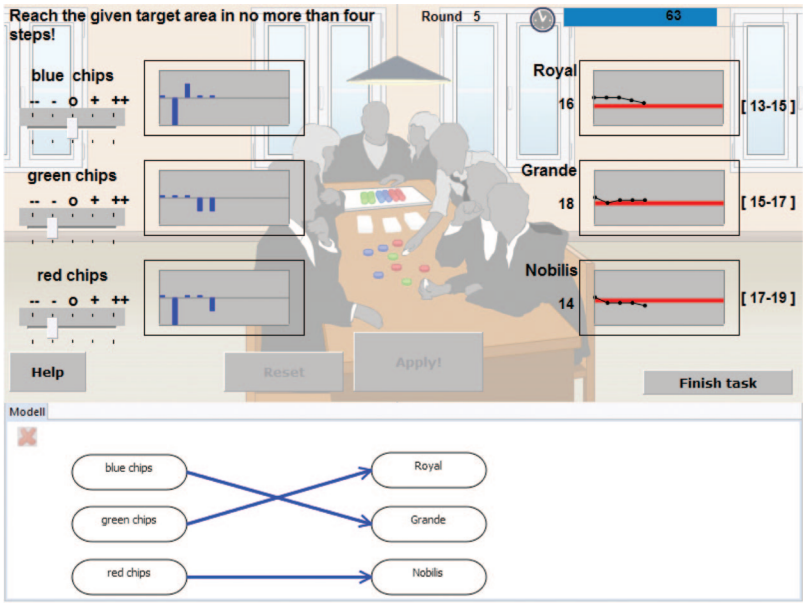

Figure 2. Screenshot of the MicroDYN task "Game Night." The controllers of the input variables range from "− −" (value = −2) to "+ +" (value = +2). The current values and the target values of the output variables are displayed numerically (e.g., current value for Royal: 16; target values: 13–15) and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

knowledge acquisition); (2) the structural stability of the CPS construct across different grade levels of high school students ages 11 to 17 years; (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,* grade point average (GPA), and parental education across these groups to assess construct validity.

With regard to (1) 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 two-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 two-dimensional model was expected to show the best fit and to fit significantly better than a one-dimensional model with the two processes combined under one first-order factor.

The structural stability of CPS, (2), 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 are between-group differences of grade levels unambiguous and able to 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 (F. F. 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 (a) that *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 (b) that CPS would predict GPA beyond *g,* as indicated by conceptual considerations and previous research. Furthermore, we expected (c) that 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 policymakers, 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 more representative sample ($N > 4{,}000$) from a study conducted in Hungary. Participants were randomly drawn from Grades 5 to 11 in Hungarian elementary schools (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 (mean age = 14.11 years, $SD = 1.83$; 46% boys). However, all analyses including those involving *g* (Hypotheses 4a and 4b) were based on a smaller subsample of students who completed both tests of CPS and *g* ($N = 486$; mean age = 14.36 years, $SD = 1.16$; 45% boys). 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* (Weiβ, 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 (Weiβ, 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; Weiβ, 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_{mother} = 3.85, SD = 1.09; M_{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, Grades 7 and 8, as well as Grades 10 and 11, to Grade Levels 5/6, 7/8, and 10/11, respectively, 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; S. S. 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 (Hypotheses 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. The two-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 (CPI) 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 two-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 one-dimensional model with all indicators combined under one general factor; however, the fit indices decreased considerably. In order to compare the two-dimensional and one-dimensional models, χ^2 values in Table 1 cannot be directly subtracted to compare them because computing the differences of χ^2 values and *df*s 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 two-dimensional model fit significantly better than the one-dimensional model (χ^2 = 86.121, df = 1, $p < .001$). After this, the two-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 (see Table 1).

In summary, the two-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 (β_1 = .82, β_1 = .78, β_1 = .80, β_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; χ^2 = .717; df = 2; $p > .05$; CFI = .999; TLI = .999;

Table 1
Goodness of Fit Indices for Testing Dimensionality of MicroDYN, Overall and by Grade Level

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

Note. χ^2 and *df* were estimated by the weighted least squares mean and variance adjusted estimator (WLSMV). CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation.

RMSEA = .001), as well as for the different grade levels (CFIs = .991–.999; TLIs = .991–.999; RMSEAs = .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 two-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 residual variances were fixed to one in all groups (as recommended by Muthén & Muthén, 2010, p. 434) instead of freely estimating residuals as is done with continuous outcomes. 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; see 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 (Grade Level 10/11 = Grade Level 7/8 $<$ Grade Level 5/6 $<$ Grade Level 9). These findings raised some concerns about the pattern of results for Grade 9; these are 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 as 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 (χ^2 = 25.546, *df* = 26; CFI = .999, TLI = .999, RMSEA = .001) and configural invariance (χ^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. 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 $<$ Grade Level 5/6 $<$ Grade Level 7/8 = Grade Level 10/11). Comparisons between the latent means of the measures of knowledge application scores showed that, once again, Grade Level 9 performed the worst, followed by Grade Levels 5/6 and 10/11, neither of which differed

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

Table 2
Goodness of Fit Indices for Measurement Invariance of MicroDYN

Model	χ^2	*df*	Compared with	$\Delta\chi^2$	Δdf	*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. χ^2 and *df* were estimated by the weighted least squares mean and variance adjusted estimator (WLSMV). $\Delta\chi^2$ and Δdf were estimated by the Difference Test procedure in MPlus (see Muthén & Muthén, 2010). Chi-square differences between models cannot be compared by subtracting χ^2s and *df*s if WLSMV estimators are used. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation.

significantly from Grade Level 9. Grade Level 7/8 performed better than all other grade levels (rank order: Grade Level 9 < Grade Level 5/6 = Grade Level 10/11 < Grade Level 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 < Grade Level 7/8 = Grade Level 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 *g* (Hypothesis 4a) used models with latent variables within structural equation modeling. However, results for CPS, *g*, 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. The 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. Although the CFT significantly predicted performance for both dimensions, the residuals of measures of knowledge acquisition and knowledge application were still highly correlated (*r*s = .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 (CFIs = .932–.992, TLIs = .960–.994, RMSEAs = .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

Table 3
Latent Mean Comparisons of Knowledge Acquisition and Knowledge Application (MicroDYN) Between Different Grade Levels

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

Note. Statistical significance of the differences between all groups was determined by *z* statistics.

Table 4
Prediction of Performance in Knowledge Acquisition and Knowledge Application (MicroDYN) by g, Overall and by Grade Level

Model	Path coefficient		R^2		Residual correlation acquisition/application	*n*
	Acquisition	Application	Acquisition	Application		
Overall	.47*** (.04)	.40*** (.05)	.22*** (.04)	.16*** (.04)	.63*** (.05)	486
Grade Level 7/8	.48*** (.05)	.39*** (.07)	.23*** (.05)	.15** (.05)	.60*** (.06)	284
Grade Level 9	.62*** (.12)	.62*** (.12)	.38** (.14)	.38** (.15)	.30*** (.10)	79
Grade Level 10	.34*** (.10)	.32*** (.11)	.11* (.07)	.11* (.07)	.62*** (.08)	123

Note. Standard errors are in parentheses.
* $p < .05$. ** $p < .01$. *** $p < .001$.

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, we thought it possible that 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, separately) 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).

Table 5
Prediction of GPA by MicroDYN and CFT

Grade	R^2 in GPA			ΔR^2	*n*
	MicroDYN	CFT	MicroDYN and CFT		
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. Significant ΔR^2s indicate significant path coefficients of CPS contributing to R^2. GPA = grade point average; CFT = Culture Fair Test 20-R; CPS = complex problem solving.
* $p < .05$. *** $p < .001$.

Results displayed in Table 5 show 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 (RMSEAs = .000–.001, CFIs = .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 hypothesized that parental education would predict performance for MicroDYN and the CFT (Hypothesis 4c). We used path analysis because of 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 performance 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 two-dimensional construct with the dimensions 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 Grade Level 7/8) with increasing grade level. However, this was not true for students in Grade 9, who performed the lowest on both dimensions, as we discuss 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.

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.

Measurement Invariance Across Grade Levels (Structural Stability)

Comparing CPS scores between grade levels requires that the assessment instrument, MicroDYN, measure 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 in Hungarian students across Grade Levels 5 through 11, which generally speaks well for the MicroDYN measure and its adoption in Hungary. 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, 2006). 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 Level 5/6 (youngest age: 11 years) to Grade Level 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.

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 the 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 was that 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 Level 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., S. S. Smith, 2006). In the specific case of Hungary, Molnár and Csapó (2007) also reported a general decrease in test scores in Grade 9 for Hungarian students, thus showing that this performance decrease is not limited to our sample. These drops in academic performance tend to recover to their pretransitional levels 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 (S. S. 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 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 we excluded Grade Level 9, a more consistent picture of latent means could be drawn. First, scores increased significantly from Grade Level 5/6 to Grade Level 7/8 for both CPS processes and *g,* showing a combined effect of school and out-of-school experiences, and even the 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 Grade Level 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 the work of Byrnes (2001), who claimed, without having studies including CPS available for his review, that both declarative knowledge and procedural knowledge increase with age during the adolescent period. Further, of the two CPS processes, the performance decrease in knowledge application from Grade Level 7/8 to Grade Level 10/11 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 Level 7/8 to Grade Level 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 life span, 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.

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 among 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, which 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 we 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 did 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

[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).

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 not surprising insofar 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 is 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, and these, paired with potentially influential transition effects, led 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 that research on CPS suffers from is 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 implicit goal 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 than 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 Z. 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 (Z. Chen & Klahr, 1999).

In line with Z. 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 those of 30 years ago, and even though the concept of life-long learning, which extends the educational span to a lifetime, has become increasingly popular (M. C. 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 (1904) 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 & 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 ed.). Washington, DC: AERA.

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): What does the WAIS–IV measure? *Psychological Assessment, 22*(1), 121–130. doi:10.1037/a0017767

Binet, A., & Simon, T. (1904). 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. doi:10.3406/psy.1904.3675

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. doi:10.1207/s15328007sem1302_7

Byrnes, J. P. (2001). *Minds, brains, and education*. New York, NY: 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, the Netherlands: Pergamon.

Chen, F. F. (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,* 1005–1018. doi:10.1037/a0013193

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. doi:10.1111/1467-8624.00081

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. doi:10.1207/S15328007SEM0902_5

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. doi:10.1016/j.intell.2011.06.004

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. doi:10.1037/0893-3200.19.2.294

Deary, I. J., Whalley, L. J., & Crawford, J. R. (2004). An "instantaneous" estimate of a lifetime's cognitive change. *Intelligence, 32,* 113–119. doi:10.1016/j.intell.2003.06.001

Diaz, L., & Heining-Boynton, A. L. (1995). Multiple intelligences, multiculturalism, and the teaching of culture. *International Journal of Educational Research, 23,* 607–617. doi:10.1016/0883-0355(96)80440-X

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 & Reasoning, 7*(1), 69–89. doi:10.1080/13546780042000046

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing, 11,* 133–142. doi:10.1007/s10339-009-0345-0

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The Euro-

pean perspective—10 years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). New York, NY: 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*(7), S55–S79. doi:10.1002/acp.2350090706

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* (2nd ed., Vol. 2, pp. 589–672). Hillsdale, NJ: Erlbaum.

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

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new measurement perspective. *Applied Psychological Measurement, 36*(3), 189–213. doi:10.1177/0146621612439620

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179–203. doi:10.1016/0160-2896(84)90008-4

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, NY: 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. doi:10.1080/10705519909540118

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.

Kihlstrom, J. F., & Cantor, N. (2011). Social intelligence. In R. J. Sternberg & S. C. Barry (Eds.), *The Cambridge handbook of intelligence* (pp. 564–581). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511977244.029

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*(4), 347–368. doi:10.1016/j.intell.2005.03.002

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 City, 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. doi:10.1207/S15328007SEM0902_1

Lynch, M., & Macbeth, D. (1998). Demonstrating physics lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 269–297). Mahwah, NJ: 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: 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. doi:10.1016/j.intell.2008.08.004

Molnár, G., & Csapó, B. (2007, August 28–September 1). *Constructing complex problem solving competency scales by IRT models using data of different age groups*. Abstract submitted at the 12th Biennial EARLI Conference, Budapest, Hungary.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Myrberg, E., & Rosen, E. (2008). A path model with mediating factors of parents' education on students' reading achievement in seven countries. *Educational Research and Evaluation, 14*(6), 507–520. doi:10.1080/13803610802576742

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, Hungary: Ecostat.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101. doi:10.1037/0003-066X.51.2.77

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

Novick, L. R., Hurley, S. M., & Francis, M. (1999). Evidence for abstract, schematic knowledge of three spatial diagram representations. *Memory & Cognition, 27,* 288–308. doi:10.3758/BF03211413

Organisation for Economic Co-operation and Development. (2004). *Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003*. Paris, France: OECD.

Organisation for Economic Co-operation and Development. (2009). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics, and science* (Vol. 1). Paris, France: OECD.

Organisation for Economic Co-operation and Development. (2010). *PISA 2012 problem solving framework* [draft for field trial]. Paris, France: 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. C. (1962). *Advanced progressive matrices, Set II*. London, England: H. K. Lewis.

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

Ree, H. M., & Carretta, T. R. (2002). g2K. *Human Performance, 15,* 3–24.

Reeve, C. L. , & Hakel, M. D. (2002). Asking the right questions about *g*. *Human Performance, 15,* 47–74. doi: 10.1080/08959285.2002.9668083

Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence, 30,* 463–480. doi:10.1016/S0160-2896(02)00121-6

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. doi:10.1016/j.lindif.2010.07.002

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology, 40*(6), 451–475. doi:10.1016/S0022-4405(02)00124-3

Salthouse, T. A., & Davis, H. P. (2006). Organization of cognitive abilities and neuropsychological variables across the lifespan. *Developmental Review, 26,* 31–54. doi:10.1016/j.dr.2005.09.001

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: Erlbaum.

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

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., . . . 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. (2000). The holey grail of general intelligence. *Science, 289,* 399–401. doi:10.1126/science.289.5478.399

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. (Original work published 1984)

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement: Issues and Practice, 14*(3), 29–35. doi:10.1111/j.1745-3992.1995.tb00865.x

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. doi:10.1521/scpq.19.1.72.29409

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, Germany: 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. 87–113). Mahwah, NJ: Erlbaum.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence, 40,* 1–14. doi:10.1016/j.intell.2011.11.003

Appendix

MicroDYN Item Characteristics and Linear Structural Equations

Item and linear structural equations	System size	Effects
Item 1	2 × 2 system	Only direct
$X_{t+1} = 1^*X_t + 0^*A_t + 2^*B_t$		
$Y_{t+1} = 1^*Y_t + 0^*A_t + 2^*B_t$		
Item 2	2 × 3 system	Only direct
$X_{t+1} = 1^*X_t + 2^*A_t + 2^*B_t + 0^*C_t$		
$Y_{t+1} = 1^*Y_t + 0^*A_t + 0^*B_t + 2^*C_t$		
Item 3	3 × 3 system	Only direct
$X_{t+1} = 1^*X_t + 0^*A_t + 2^*B_t + 0^*C_t$		
$Y_{t+1} = 1^*Y_t + 2^*A_t + 0^*B_t + 0^*C_t$		
$Z_{t+1} = 1^*Z_t + 0^*A_t + 0^*B_t + 2^*C_t$		
Item 4	3 × 3 system	Only direct
$X_{t+1} = 1^*X_t + 2^*A_t + 0^*B_t + 0^*C_t$		
$Y_{t+1} = 1^*Y_t + 0^*A_t + 2^*B_t + 2^*C_t$		
$Z_{t+1} = 1^*Z_t + 0^*A_t + 0^*B_t + 2^*C_t$		
Item 5	3 × 3 system	Only direct
$X_{t+1} = 1^*X_t + 2^*A_t + 2^*B_t + 0^*C_t$		
$Y_{t+1} = 1^*Y_t + 0^*A_t + 2^*B_t + 0^*C_t$		
$Z_{t+1} = 1^*Z_t + 0^*A_t + 0^*B_t + 2^*C_t$		
Item 6	2 × 3 system	Direct and indirect
$X_{t+1} = 1.33^*X_t + 2^*A_t + 0^*B_t + 0^*C_t$		
$Y_{t+1} = 1^*Y_t + 0^*A_t + 0^*B_t + 2^*C_t$		
Item 7	3 × 3 system	Direct and indirect
$X_{t+1} = 1^*X_t + 2^*A_t + 0^*B_t + 0^*C_t$		
$Y_{t+1} = 1^*Y_t + 2^*A_t + 0^*B_t + 0^*C_t$		
$Z_{t+1} = 1.33^*Z_t + 0^*A_t + 0^*B_t + 2^*C_t$		

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

Received December 16, 2011
Revision received November 5, 2012
Accepted December 26, 2012 ■