PhD-FLSHASE-2016-26
The Faculty of Language and Literature, Humanities, Arts and Education

DISSERTATION

Defence held on 14/10/2016 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PSYCHOLOGIE

by

Matthias STADLER
Born on 08 July 1987 in Munich (Germany)

COMPLEX PROBLEM SOLVING IN UNIVERSITY SELECTION

Dissertation defence committee

Dr Samuel Greiff, dissertation supervisor
Associate-Professor, Université du Luxembourg

Dr Frank M. Spinath
Professor, Universität des Saarlandes

Dr Romain Martin, Chairman
Professor, Université du Luxembourg

Dr Stephan Kröner
Professor, FAU Nürnberg

Dr Christoph Niepel, Vice Chairman
Université du Luxembourg

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Samuel Greiff for his valuable support during the past three years. His unsparing but always very constructive feedback helped me to improve both my scientific thinking as well as my writing. I highly appreciate all the experiences he made possible for me such as my trips to Princeton and Beijing, all the conferences, writing grant proposals, and guest editing for a Journal. And last but not least, he taught me not to save my chocolate for later.

My sincere thanks also go to Prof. Dr. Romain Martin and Prof. Dr. Frank Spinath for the additional support in their function as my CET. This macroscopic view on the progress of my thesis helped me not to lose track over the daily hassles of research life.

Furthermore, I would like to thank Dr. Christoph Niepel for his inspiring daily supervision both at the office and during various long car rides. The discussions we had during endless traffic jams lead to some of the most interesting ideas of my dissertation.

Dr. Sascha Wüstenberg helped me find my way into the world of complex problem solving. His enormous knowledge and our resulting discussions about research, statistics, Magic, and Star Trek were the highlights of most of my days at work and filled various great evenings.

I want to express my gratitude to my friends and colleagues Dr. Nicolas Becker, Julia Rudolph, Jakob Mainert, Jonas Neubert, Dr. André Kretzschmar, Katarina Krkovic, Jan Dörendahl, and all remaining and former members of the CPS team. It was great to have you around during my time as a doctoral student.

Finally, I am deeply grateful to my family for their continuous help and support through the whole time. Most of all, I would like to thank my wife Ina who was never tired of hearing my crazy ideas. Without her love, help, creativity, intelligence, and patience I would not be where I am now.

Summary

This thesis investigates the utility of complex problem solving (CPS) in the prediction of university success. Previous research focused mainly on the relation between CPS and primary or high-school success, ignoring that the demands at university are actually far more complex than at lower school forms. On the other hand, CPS has often claimed to be redundant to intelligence, a well-established predictor of university success. This thesis, therefore, attempts to answer four complementary research questions dealing with (1) innovations in the assessment of CPS (2) the relation between CPS and intelligence in the prediction of university success (3) the relation between CPS and university success and (4) the defining characteristics of CPS tasks.

By applying a vast array of different methods ranging from theoretical suggestions on the improvements of CPS assessment to meta-analyses, structural equation modeling, and item response models, this work is therefore the first extensive investigation of the utility of CPS in the prediction of university success and considerably extends the research on the validity of CPS as a construct. More specifically, this work introduces (1) the theoretical foundation for a multiple task approach to measure CPS that is in many ways superior to previous measurement approaches and thus allows for a reliable assessment of individual CPS skills. This is followed by (2) a meta-analytic investigation of the empirical relation between CPS and intelligence, an established predictor of university success, to rule out empirical redundancy between the two related constructs. Given the often reported strong but far from perfect relation between CPS and intelligence this work further describes (3) the investigation of incremental validity of CPS over and above intelligence in the prediction of university success. Finally, in order to be used as potential tools in

university selection, CPS tasks need to be easy to create and adapt. This work is therefore concluded by (4) an analysis of the difficulty of CPS tasks as a function of defining characteristics laying the ground for the efficient generation of new tasks. In summary, the present work addresses several important gaps in existing research both on CPS assessment and the prediction of university success.

In Chapter 1, the societal need, as well as the theoretical and empirical foundation for this research, are introduced. This is followed by a brief description of the four empirical papers that are the main body of this thesis. The full papers are located in Chapters 2 to 5. Papers 1, 2, and 4 are already published after having successfully passed peer-review, Paper 3 is currently under review. Chapter 1 and Chapter 6 refer to additional papers including supplementary contributions of the author of this thesis on both CPS and the prediction of university success, which are listed as “additional papers” on page 9.

The first empirical paper included in this thesis introduces the theoretical foundations of the Multiple Complex Systems (MCS) approach to assess CPS skills. Other than previous CPS assessment approaches, that used single very large and complex microworlds, the MCS approach relies on multiple smaller microworlds that are combined into one assessment instrument. This innovation leads to several important advantages over previous approaches including a highly increased reliability, variations in item difficulty, and scalability that allows for the application of advanced statistical models (cf. Chapter 2).

The second paper meta-analytically investigated the relation between CPS and intelligence. The main finding was that the two constructs were highly related but not redundant to each other. Furthermore, this relation was moderated by the approach used to assess CPS with only small average correlations between classical measures

of CPS and intelligence but rather strong correlations for MCS measures of CPS (cf. Chapter 3).

The third contribution investigated the validity of CPS in the prediction of university students' objective and subjective academic success. Based on two independent samples it could be shown that CPS was substantially related to students' grade point average (GPA) and subjective ratings of their success. This effect remained significant even after intelligence was controlled for (cf. Chapter 4).

The fourth and final paper investigated the difficulty of MCS tasks. Based on six basic characteristics, it was possible to predict the tasks' difficulty almost perfectly demonstrating a deep understanding of the defining aspects of MCS tasks. This lays the foundation for the efficient or even automatic generation of new tasks with known qualities (cf. Chapter 5) as is necessary for any form of high-stakes selection.

Chapter 6 provides a general discussion of this research and its implications. Taken together, all four papers support the potential use of CPS in university selection. CPS can be assessed reliably, is not redundant to already established predictors, and incrementally explains students' individual differences in academic performance. Finally, CPS tasks are well understood and can be easily created or adjusted making them very well suited for high-stakes assessment.

After this summary of results, strengths of the papers are outlined and shortcomings combined with an outlook for future research are discussed. In summary, this thesis advances knowledge about CPS and emphasizes its usefulness as an additional predictor of university success.

Contents

Publication list for this cumulative dissertation

Paper 1:

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2014). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*, 356-382.

Paper 2:

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92-101.

Paper 3:

Stadler, M., Becker, N., Schult, J., Niepel, C., Spinath, F. M., Sparfeldt, J. R., & Greiff, S. (submitted). The logic of success: The relation between complex problem solving skills and university achievement. *Higher Education.*

Paper 4:

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, *65*, 100-106.

Additional Papers

Stadler, M., Aust, M., Becker, N., Niepel, C., & Greiff, S. (2016). Choosing between what you want now and what you want most: Self-control explains academic achievement beyond cognitive ability. *Personality and Individual Differences, 94*, 168-172.

Stadler, M., Becker, N., Greiff, S. & Spinath, F. M. (2015). The Complex Route to Success: Complex Problem Solving predicts University Success. *Higher Education Research and Development, 35*, 365-379.

The research described in this paper was part of my Diploma Thesis

Becker, N., Stadler, M., & Greiff, S. (2015). Noch mehr als Intelligenz? Komplexes Problemlösen im alltäglichen Leben [More than intelligence? Complex problem-solving in everyday live]. *InMind, 6*.

Greiff, S., Stadler, M., Sonnleitner, P., Wolf, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence, 50*, 100-113.

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning, 19*, 127-146.

1

Introduction

The only person who is educated is the one

who has learned how to learn and change.

Carl Rogers (1902 - 1984)

1.1 Introduction

Education plays a critical role in fostering social progress. Correspondingly, it is a sign of lasting societal change that the access to education continues to expand in Organization for Economic Cooperation and Development (OECD) countries (OECD, 2014). This change in societies over only a couple of generations, from a time when only an elite few were highly educated to today's situation where three-quarters of the population have at least an upper secondary education, is one whose consequences are still unfolding. Almost 40% of 25-34 year-olds nowadays have a tertiary education, a proportion that is 15 percentage points larger than that of 55–64 year-olds; and in many countries, this difference exceeds 20 percentage points (OECD, 2014). This great achievement comes with new challenges. Providing higher education for such large proportions of a population represents a massive investment by individuals, organizations, and societies that need to be economically justifiable (Walker & Zhu, 2003). Among the member countries of the OECD, an average of 6.2% of the gross domestic product is spent on educational activities, and the average young person in these countries will receive an education until the age of 22 (OECD, 2007). Education still remains one of the best individual and societal investments (Elias, & Purcell, 2004) but only if the students actually complete their degree successfully.

On the other hand, not all university programs are equally famous with students. While some programs such as philosophy appear less attractive, others such

as psychology or medicine have far more applicants than they could possibly handle (Statista, 2016).

The search for fair and feasible selection procedures of adequate applicants and valid prediction of potential success at university, therefore, becomes increasingly important. Predictors of university success have been researched for over a century (Bingham, 1917), finding that adding information about psychological constructs to the information gained from previous academic achievements such as high school grade point average (GPA), increases the accuracy of predicting university success (Kuncel, Hezlett, & Ones, 2001). Most notably, individual differences in intelligence have consistently been found to add value in explaining the variation within university success (Richardson, Abraham & Bond, 2012).

Measures of intelligence do not provide detailed information about a person's skills of acquiring and applying new knowledge about a dynamic problem or system, tough (Wüstenberg, Greiff & Funke, 2012). Measures of complex problem solving (CPS) on the other hand, aim to assess these skills, which are vital for the successful completion of every higher degree in today's society (Koeppen, Hartig, Klieme & Leutner, 2008). As opposed to intelligence, CPS requires participants to actively interact with and explore new and partly opaque problem situations in order to gather the information necessary to find solutions to the problems. For that purpose, participants are given simulated microworlds they can manipulate receiving feedback on the effects of their actions. However, even though several studies provide strong evidence in favor of the validity of CPS in the prediction of success in various environments such as school success (e.g., Schweizer, Wüstenberg & Greiff, 2013) or job performance (Danner et al., 2011), no extensive research on the introduction of tests of CPS into university applicant selection has been conducted so far.

This thesis will therefore comprehensively address the question on the validity of CPS as a construct and its utility in university selection. In that, the thesis will answer four complementary research questions:

Research Question 1: Can CPS be reliably measured without the influence of previously acquired knowledge?

Research Question 2: Is CPS theoretically or empirically redundant to intelligence?

Research Question 3: Are measures of CPS valid in predicting indicators of university success? If so, do they show incremental validity over and above measures of intelligence?

Research Question 4: Is it possible to predict the difficulty of CPS measures based on their defining characteristics?

Each of the four papers primarily addresses one of these research questions and they are ordered accordingly (Chapters 2-5). Prior to that, a short description of the understanding of university success underlying this thesis and previous research on its prediction is given in section 1.2. This is followed by an introduction to the construct of CPS as well as to measurement approaches to assess individual differences in CPS competency in Section 1.3. Preceding results on the relation between CPS and university success are reported in Section 1.4, followed by a brief description of the four individual papers in section 1.5.

1.2 University success

1.2.1 Definition

The overall aim of this thesis was to investigate the validity of CPS as a construct and to examine its utility in the prediction of university success.

Unfortunately, there is no generally accepted understanding on the definition of university success. Most attempts to define what constitutes success at university therefore become tautological, by defining university success based on one or more criteria of university success (see Anderson, 2003). In that, the definition of university success highly depends on the person, the peer group, or the institution in question (Konegen-Greiner, 2001). Especially on the institutional side (i.e., universities or governments) definitions tend to use very narrow or "hard" criteria such as focusing merely on students' GPA (e.g., Wissenschaftsrat, 2004). Others, mostly on the personal side (i.e., students and lecturers), also include "soft" criteria such as personal satisfaction or other more subjective estimates of students' success (Lattner, & Haddou, 2013). The notion of "university success" can thus have various different meanings for both students and institutions, such as graduating with a high GPA, graduating as fast as possible, learning as much as possible, completing the degree, prospects on future job success, or the subjective satisfaction with the degree (Kunina, Wilhelm, Formazin, Jonkmann, & Schroeders, 2007). A graphical representation of this distinction is provided in Figure 1.

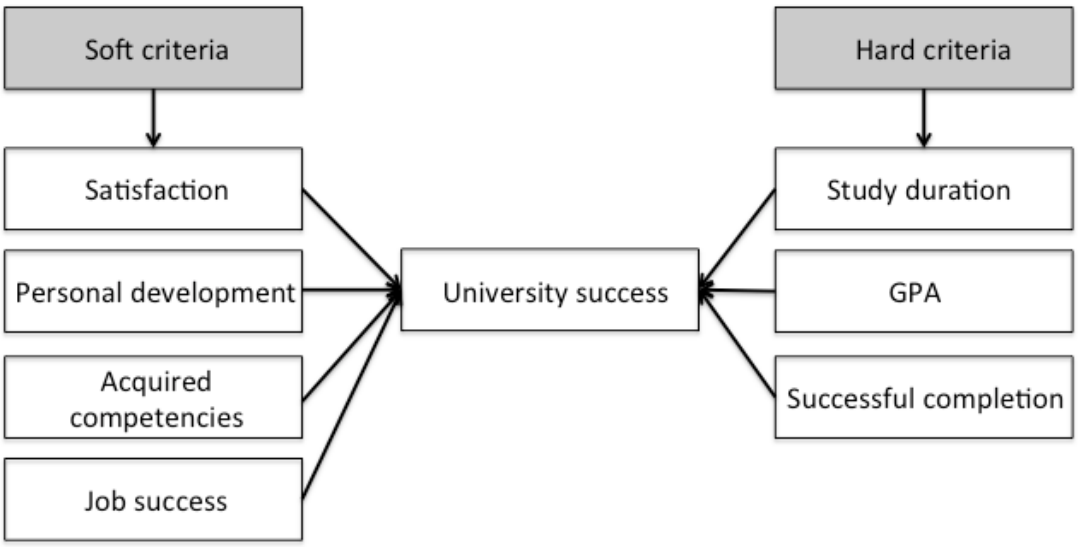

Figure 1. Criteria of university success (adapted from Lattner, & Haddou, 2013)

The “hard” criterion of students’ GPA is by far the most widely used and studied measure in tertiary education (Bacon & Bean, 2006; Richardson et al., 2012). Students’ GPA is the most salient criterion for students, is economically available, and correlates strongly with variables of interest to educational researchers such as intelligence, motivational strategies or certain personality traits (Richardson et al., 2012). GPA has been found to be a key criterion for postgraduate selection and employment and thus represents a valid predictor of socioeconomic success (Strenze, 2007).

The sole use of GPA as an indicator of university success has often been criticized, though. Johnson (2003) for example, called grade inflation (very good or excellent grades becoming more and more commonplace) a crisis in university education. He further argues that every university uses multiple and sometimes very different grading approaches to evaluate students (see also Babcock, 2010). These grading disparities between universities, study programs, and even between different examiners at the very same program, as well as the aspect of grade inflation impair a fair and reliable assessment of students’ competencies. This has serious consequences on student’s future perspectives concerning the chance of finishing university with a higher GPA.

Thus, GPA has, despite its considerable advantages, some noteworthy limitations as widespread indicator of students’ university success.

Universities and researchers alike have responded to these limitations by including “soft” criteria into their definitions of university success as well. Most importantly, the subjective value that students attribute to specific indicators of university success may vary from student to student. In other words, students may, for example, consider a passing grade as either success or failure depending on their

subjective expectations. Lattner and Haddou (2013) conducted an interview study with students from all faculties resulting in a total of 10 subcategories of university success. While grades and the successful completion of the program were important to most students (though not all), softer aspects such as individual progress, practical relevance, fun, and reaching individual goals were considered equally important. In order to fully capture the heterogeneity of the construct of university success, "hard" and "soft" criteria of success should thus be considered complementary (Duckworth, Weir, Tsukayama, & Kwok, 2012).

Within this thesis, students' university success will therefore be considered a multidimensional construct consisting of both "hard" and "soft" criteria, which should be considered complementary in order to gain a holistic understanding of the construct.

1.2.2 Predictors of university success

Based on the multidimensional conception of university success underlying this thesis, managing a university program requires dealing with a complex system of academic tasks. These may include new learning and study behaviors, scientific thinking, social obligations, and various other demands that are either unique to university studies or at least more important than in high school (Parker, Summerfeldt, Hogan, & Majeski, 2004). It is therefore no surprise that numerous factors have been suggested to influence students' university success, such as cognitive (e.g., intelligence or previous academic achievement; e.g., Formazin et al., 2011), noncognitive (e.g., motivational factors, self-regulatory learning strategies, personality traits, students' approaches to learning, or psychosocial contextual influences; for an overview see Richardson et al., 2012), and demographic (e.g., age or socio demographic background; e.g., Robbins et al., 2004).

The main focus of this thesis is placed on the cognitive predictors of university success. Most importantly, intelligence has been established as one of the strongest and most reliable predictors of academic achievement since the early 20th century explaining about 25% of the variance in university students' GPA (e.g., Binet & Simon, 1916; Bingham, 1917; Jensen, 1998; Kuncel, Hezlett, & Ones, 2004). This means however, that equally intelligent students may differ largely in their university success. Other cognitive abilities have therefore come into the focus of researchers recently. Especially in tertiary education, where student selection procedures reduce variation in intelligence scores, the predictive value of intelligence is limited (Furnham, Chamorro-Premuzic & McDougall, 2003). Highly selective academic institutions show only very low variation in intelligence among their students (Jensen, 1998). Other cognitive skills than intelligence may consequently add important incremental information to the accurate prediction of performance at university level.

This becomes particularly evident in the differential development and prediction of "hard" and "soft" indicators of university success (e.g., Harackiewicz, Barron, Tauer, & Elliot, 2002). Whereas intelligence consistently predicts university students GPA, subjective or "soft" indicators of university success seem to be more closely linked to psychosocial and study skill factors (Robbins et al., 2006). For instance, Robbins and colleagues (2004) investigated the role of study skill factors as predictors of university outcomes in addition to other well-established cognitive predictors. Their meta-analysis showed that academia-related skills, defined as "cognitive, behavioral, and affective tools and abilities necessary to successfully complete task, achieve goals, and manage academic demands" (Robbins et al., 2004; p. 267), to be meaningful predictors of both university GPA ($r = .13$) and university retention rates ($r = .30$).

The aim of this thesis was therefore to investigate whether one such academia-related skill, namely CPS, would be useful as an addition to intelligence in the prediction of university success. The following section will define the construct of CPS as well as introduce its assessment and the distinction between CPS and intelligence.

1.3 Complex problem solving (CPS)

1.3.1 Definition of CPS

Imagine a university student who just started their first years as a freshman. In order to deal with the new demands of university life, the students need to generate knowledge about the universities inner workings such as choosing which lectures to attend, when to write exams, or learning how to borrow books from the library. In addition, they need to acquire new study habits adapted to the university requirements with its complex content. After having explored the university for a while, they will be able to apply that generated knowledge in order to succeed in their programs. This is a typical situation considered as a complex problem involving dynamic interaction with a yet unknown system.

Complex problems contain multiple variables (complexity) that are interrelated (connectivity) and may change either as a result of the problem solvers manipulations or over time (dynamics). The problems' structure is partially or fully opaque to participants (intransparency) and needs to be actively explored. This is summarized in Buchner's definition of CPS as:

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

successful exploration and integration of the information gained in that process."

(Buchner as cited by Frensch & Funke, 1995; p. 14)

As described in the definition, such complex problems have no obvious method of solution and barriers between the initial state (e.g., having to choose the most appropriate lectures) and the goal state (e.g., achieving good grades) have to be reduced by applying non-routine cognitive activities (Funke, 2012; Mayer, 1992; Mayer & Wittrock, 2006). Problem solvers dealing with such complex problems face two main demands: generating knowledge about the systems' structure (i.e., knowledge acquisition; Novick & Bassok, 2005) and the need to reach a certain goal by applying knowledge gathered beforehand (i.e., knowledge application; Novick & Bassok, 2005). While acquiring knowledge in complex problems, problem solvers build a problem representation and derive a problem solution, which are the two major components of the problem solving process accountable to all kinds of problem solving (Mayer, 2003; Mayer & Wittrock, 2006; Novick & Bassok, 2005).

1.3.2 Assessment of CPS

To allow for an active interaction between the student and the assessment instrument, the assessment of CPS necessarily requires a computer-based assessment (Frensch & Funke, 1995). With the advancement of computer technology, various CPS tasks have therefore evolved following different approaches. The first computer-based CPS tasks were developed in the early 1980s. The aim was to administer task environments with a high resemblance to the real world and the goal of producing a reliable and ecological valid measure of CPS that sufficiently emulated real world problems. The complex problem "Lohausen" (Dörner, Kreuzig, Reither, & Stäudel, 1983), for instance, required a participant to govern a small city. This city was intricately modeled with over 1000 separate interconnected variables. Such

classical measures of CPS had a high level of face validity, as they seemed to mirror real life problem solving. Their psychometric properties however were insufficient (Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015). Unsatisfactory reliability and validity raised doubt on the measurability and validity of the construct of CPS itself (Kröner et al., 2005; Wüstenberg et al., 2012). Moreover, knowledge about the real world situation emulated by the classical measure of CPS strongly influenced performance in these tasks. This limited the usability of classical measures of CPS such as Lohhausen as assessment instruments. Funke (2001) responded to these problems by introducing Linear Structural Equation systems (LSE) and Finite State Automata (FSA) as formal frameworks that allow for the description of underlying task structures. Both of these frameworks enabled the creation of single complex systems, which are independent of any semantic embedment (Greiff et al., 2015). These single complex systems specify an underlying system that can be applied to multiple, arbitrary semantic contexts thus removing the influence of any previous knowledge. Especially the LSE formalism has been widely adopted by CPS research leading to the development of a considerable number of single complex systems [e.g., "Multiflux" (Kröner, 2001) or "FSYS" (Wagener, 2001)]. In a further advancement Leutner, Klieme, Meyer, and Wirth (2004) used a combination of two single complex systems for measuring CPS as an aggregated score. Greiff, Wüstenberg, and Funke (2012) extended this idea for the development of the multiple complex systems (MCS) approach. The MCS approach solves several measurement issues by using multiple small tasks, rather than one single large task as in classical measures of CPS or single complex systems (Greiff et al., 2015). The first in assessment tools following the MCS approach such as MicroDYN (Greiff et al., 2012) or Genetics Lab (Sonnleitner et al., 2012) were based on LSE. Later, the approach was extended to

FSA with the development of MicroFIN (Neubert, Kretzschmar, Wüstenberg, & Greiff, 2014). These MCS measures of CPS were developed with a clear focus on quality and showed significantly higher reliability than classical measures of CPS. Figure 2 shows an example of a typical MicroDYN (Greiff et al., 2012) task. In this task a problem solver needs to first figure out the effect of three generically labeled ingredients of a perfume (Norilan, Miral, and Carumin) on three characteristics of the perfume (Sweet, Flowery, and Fruity). After the relation was explored and plotted below the task, the problem solver needs to reach specific popularity values for all three products (the red lines in the graphs on the right side of the task) in no more than four steps.

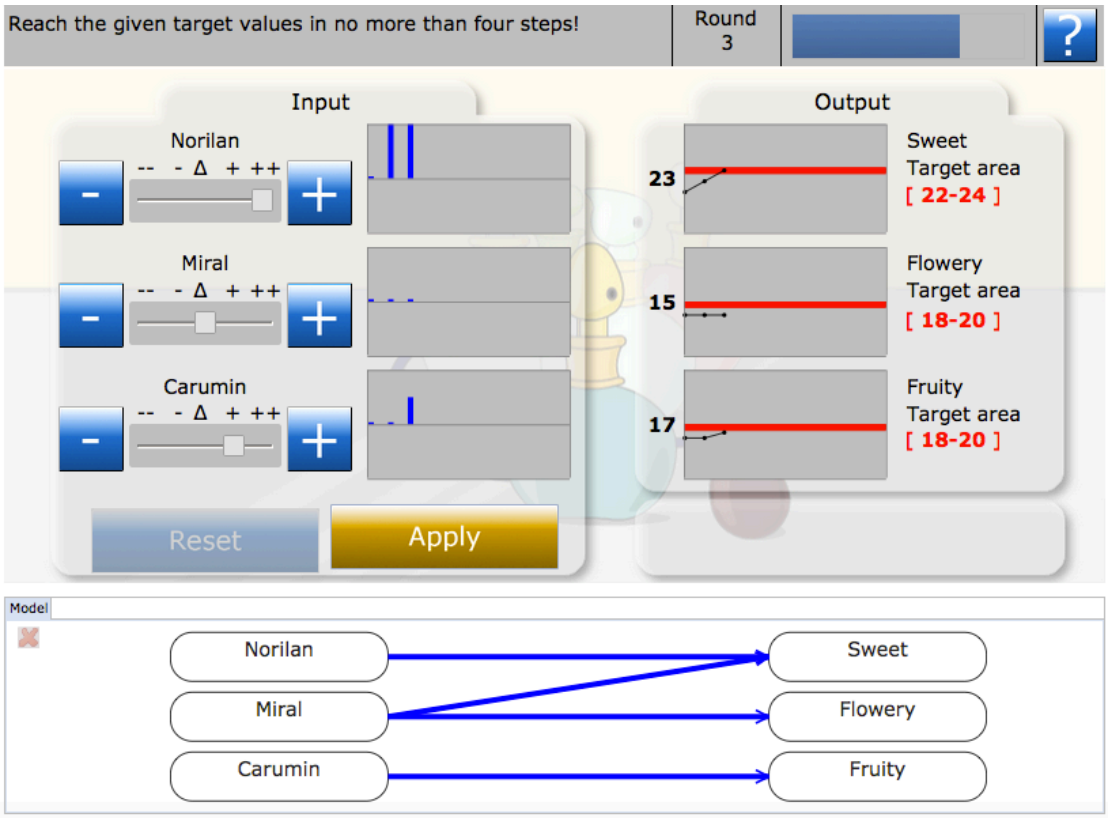

Figure 2. Screenshot of MicroDYN's graphical interface during the knowledge application phase. Horizontal lines indicate the target values for the outcomes

variables. The underlying relations between the variables are given in the lower section of the figure.

1.3.3 CPS and intelligence

Both on the conceptual basis (Funke & Frensch, 2007) and on an assessment level (e.g., Kröner et al., 2005), CPS has often been compared to intelligence. Various defining features of CPS such as the integration of information or the detection of underlying structures are part of most definitions of intelligence (Sternberg & Berg, 1986). On the other hand, the dynamic and opaque aspects of CPS are not established in the current conceptions of intelligence such as the Cattell–Horn–Caroll (CHC) theory (McGrew, 2009). These aspects of CPS may therefore be important additions for the understanding of human ability (Dörner & Kreuzig, 1983; Greiff et al., 2013).

This theoretical ambiguity is reflected in empirical findings on the relation between CPS and intelligence. Several early studies on the relation between CPS and intelligence indicated that psychological assessments of intelligence were unable to explain variance in CPS (Brehmer, 1992; Rigas & Brehmer, 1999). Kluwe,Misiak, and Haider (1991) summarized 11 of these early studies on the relation between CPS and intelligence and concluded that most of them failed to show a close relation between intelligence scores and CPS performance measures. This led several researchers to suggest CPS to be a cognitive construct mostly independent from intelligence (Putz-Osterloh, 1985). Rigas and Brehmer (1999) summarized this view in the different-demands hypothesis. This hypothesis suggests that CPS tasks demand the performance of more complex mental processes than intelligence measures do, such as the active interaction with the problem to acquire knowledge on the problem environment.

Whereas there is some support for the different-demands hypothesis (e.g., Joslyn & Hunt, 1998), more recent studies challenge it. In a comprehensive study, Gonzalez, Thomas, and Vanyukov (2005) found correlations ranging from r=.33 to r=.63 between various measures of CPS and measures of general intelligence. Similarly, Süß, Kersting, and Oberauer (1991) reported correlations of r = .40 between Tailorshop performance measures (Tailorshop being one of the most frequently used measures of CPS) and measures of general intelligence. Based on these moderate to strong correlations several researchers came to argue that measures of CPS would be almost redundant to measures of intelligence (Mayer et al., 2013; Wittmann & Süß, 1999).

An explanation for these inconsistent findings regarding the relation between CPS and intelligence may lie in the operationalization of CPS. In line with the different-demands hypothesis, the operationalizations of CPS differed in their level of complexity with classical measures being very complex and MCS measures minimally complex. Correspondingly, the relation between CPS and intelligence may differ depending on the CPS measure used. An alternative explanation for the fuzzy results of studies on the relation of intelligence and CPS could lie in the semantic embedment of CPS tasks. The Elshout–Raaheim hypothesis (Elshout, 1987; Raaheim, 1988; see also Leutner, 2002) proposes an inverted U-shaped relation between the correlation coefficient as the dependent variable and the amount of available domain-specific knowledge as the independent variable. As classical measures of CPS emulated real-world problems, domain specific knowledge could be used to solve the problems, thus limiting the relevance of individual intelligence. More recent measures of CPS such as MCS measures are less dependent of a semantic context, and, thus, less domain specific knowledge can be used. This should

result in a stronger relation between performance in modern CPS tasks and intelligence.

In summary, the relation between CPS and intelligence remains unclear. This is particularly important for this thesis, which investigates the incremental validity of CPS in predicting university success over and above intelligence.

1.4 CPS and university success

The ability to deal with dynamically changing and opaque systems should be necessary to be successful at any academic institution. Support for this notion comes from several articles reporting that CPS predicts high school grades beyond measures of intelligence (Greiff et al., 2013; Wüstenberg et al., 2012; see Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016 for divergent findings) or working memory capacity (Schweizer et al., 2013). As outlined above, the demands posed by university programs should be more complex and cognitively challenging than those encountered at high school. In her model of university success, Ferrett (2000) describes cognitive skills such as time management, preparing for and taking examinations, or using information resources as the focal point of the freshman year experience. University students face a variety of new challenges such as learning and applying study habits in a more complex academic environment and generally discovering how to function as independent and academically successful adults, which requires planning and problem-solving competencies (e.g., acquiring knowledge about new problems or prioritizing sub goals). In other words, students need to solve complex problems to be successful in college. Surprisingly though, only one study has investigated the relation between CPS and university success to date (i.e., Stadler, Becker, Greiff, & Spinath, 2015). This study, which will function as a starting point for this thesis found a substantial relation between CPS and both GPA

and subjective university success of business students (β = .38) that remained significant even after general intelligence was controlled for.

However, the study by Stadler and colleagues was severely limited in its generalizability. First, the sample size used was rather small (N = 78) and did not allow for advanced statistical analyses such as structural equation modeling. Furthermore, the sample consisted exclusively of business students and was thus rather homogeneous. Regarding the measures used, both the very broad measure of intelligence and the highly complex measure of CPS may have further influenced the results thus additionally limiting their generalizability.

The aim of this thesis will therefore be an extensive investigation of the validity of CPS as a construct and its utility in predicting students' university success. For this purpose, Paper 1 provides a comprehensive review of CPS measurement approaches introducing MCS measures and comparing them to other established measures. Paper 2 investigates the relation between CPS and intelligence and provides a meta-analysis on its dependency on different measurement approaches. Once these first two papers have determined the most adequate methods to operationalize both CPS and intelligence, Paper 3 can investigate the validity of CPS in the prediction of university success as well as its incremental value over and above intelligence. Finally, Paper 4 will further validate CPS measures as applicable in high-stakes assessments by demonstrating the good predictability of MCS tasks' difficulty.

1.5 Preview of the individual papers

1.5.1 Preview of Paper 1

Paper 1 introduces the MCS approach as a way to reliably measure individual differences in CPS. After defining the construct, the paper gives an overview over the formal frameworks for describing complex problems. These consist of LSE, which

model the relation between variables in a complex system as a set of linear structural equations, and FSA, which describe a complex system as a set of variables with a finite amount of states. Both of these frameworks were used to develop a multitude of CPS measures.

However, all of these measures consisted of one single, highly complex task. This leads to several measurement issues that occur when a test is composed of single tasks only. Specifically, test with only one single item have fixed item difficulty that cannot be adjusted to the assessment situation. Furthermore, these measures show low reliability that can hardly be determined as every action within the system strongly depends on the previous action. Along that line of thought, random errors in the early phases of the measurement can have a large impact on the final result as they influence every succeeding action.

MCS measures avoid these issues by combining multiple small CPS tasks into one measure. That way, it is possible to vary the difficulty of the measure by the combination of differently difficult tasks. The reliability of such measures can be determined by the internal consistency across the individual tasks. Correspondingly, errors in the first tasks do not necessarily influence the behavior in the following tasks and thus do not overly skew the final assessment result. MCS tasks can be based on both the LSE and the FSA frameworks.

MCS measures thus represent an important advancement in the measurement of CPS. On the other hand, MCS measures need to consist of tasks that are considerably smaller and less complex than those of CPS measures using only a single task. Thus MCS measures trade qualities such as reliability and scalability against the possibility to simulate extremely complex systems.

1.5.2 Preview of Paper 2

Paper 2 meta-analytically examines the nature and magnitude of the relation between CPS and intelligence. Theoretically, researchers have hypothesized the two constructs to be everything from completely separate (e,g,, Quelle) to identical (e,g,, Quelle). Over the course of almost four decades, empirical studies yielded results supporting both arguments with correlation coefficients from r = -.3 to r = .8. To summarize these results and search for moderating factors, the data of 47 studies containing 60 independent samples and a total sample size of 13,740 participants was collected. Across all samples, the analysis revealed a medium correlation between CPS and intelligence with an average effect size of M(g) = .433.

Additional moderator analyses investigated whether the operationalization of CPS and intelligence could explain the inconsistencies among the various studies. Whereas there were no significant differences in the correlation considering the operationalization of intelligence, the approach used to measure CPS moderated the correlation of CPS and intelligence. The MCS measures of CPS yielded the strongest associations between the two constructs. Classical measures of CPS on the other hand led to a substantially smaller correlation between CPS and intelligence.

The results thus clearly show a medium to strong relation between CPS and intelligence. On the other hand it could also show that the two constructs are far from redundant to each other.

1.5.3 Preview of Paper 3

The aim of Paper 3 was to investigate the role of CPS in undergraduate students' university success in two independent studies. In that CPS should not only predict different indicators of university success but also show incremental validity over and above intelligence. Following the findings of the Papers 1 and 2, CPS was

operationalized using an MCS measure and intelligence was assessed using a short reasoning measure. This allows for a reliable measurement of CPS (cf. Paper 1) while not underestimating the relation between CPS and intelligence. To reach a high generalizability of the findings, the research question is investigated with two independent samples.

In Study 1, university GPAs and subjective evaluation of academic success were collected for 165 university students who predominantly studied psychology. CPS made a significant contribution to the explanation of GPAs and the subjective success evaluations when controlling for intelligence.

To further investigate this effect, Study 2 relied on an independent and more heterogeneous sample of 216 university students. The findings of Study 1 were replicated in this study. Thus, the results of both studies suggest a link between individual differences in CPS and the abilities necessary to be academically successful university education.

1.5.4 Preview of Paper 4

Paper 4 further investigates the utility of CPS tasks in high-stakes assessment situations such as university applicant selection. Fairness and security aspects are of outmost importance in any applicant selection. One of the major concerns of high-stakes testing is therefore the integrity of items, which can be severely compromised by repeated use. Automatic item generation, as a means of minimizing the effort necessary to create new items, can present a cost efficient and suitable way to tackle this problem. To generate items automatically, test items must be converted into an item model that is a prototypical representation of the test items to be generated. Such prototype items model could determine the difficulty of any theoretically describable item a priori. However, no such item model exists for CPS tasks.

To fill this gap, we analyze data of 3056 Finnish students using a linear logistic test model (LLTM). The LLTM models the likelihood of solving an item correctly (i.e., the item's difficulty) as a function of individual ability and a linear combination of specific item characteristics and their relative contribution to item difficulty. Our results suggest that the difficulty of MCS tasks is almost perfectly predictable by six basic characteristics; namely, the use and number of (1) eigendynamics, the number of (2) input and (3) output variables, the number of (4) input and (5) output variables not related to any other variables, and (6) the total number of relations between all variables. In addition, we provide evidence for the necessity of differentiating between difficulty of controlling a CPS task (knowledge application) and understanding it's underlying system (knowledge acquisition).

2

Assessing complex problem solving skills with multiple complex systems

This article is available as:

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2014). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*, 356-382.

Thinking & Reasoning
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ptar20

Assessing complex problem-solving skills with multiple complex systems

Samuel Greiff[a], Andreas Fischer[b], Matthias Stadler[a] & Sascha Wüstenberg[a]

[a] University of Luxembourg, Institute of Cognitive Science and Assessment, Luxembourg, Luxembourg

[b] Department of Psychology, University of Heidelberg, Heidelberg, Germany

Published online: 20 Dec 2014.

To cite this article: Samuel Greiff, Andreas Fischer, Matthias Stadler & Sascha Wüstenberg (2014): Assessing complex problem-solving skills with multiple complex systems, Thinking & Reasoning, DOI: 10.1080/13546783.2014.989263

To link to this article: http://dx.doi.org/10.1080/13546783.2014.989263

Thinking & Reasoning, 2014
http://dx.doi.org/10.1080/13546783.2014.989263

Routledge
Taylor & Francis Group

Assessing complex problem-solving skills with multiple complex systems

Samuel Greiff[1], Andreas Fischer[2], Matthias Stadler[1], and Sascha Wüstenberg[1]

[1]University of Luxembourg, Institute of Cognitive Science and Assessment, Luxembourg, Luxembourg
[2]Department of Psychology, University of Heidelberg, Heidelberg, Germany

In this paper we propose the multiple complex systems (MCS) approach for assessing domain-general complex problem-solving (CPS) skills and its processes knowledge acquisition and knowledge application. After defining the construct and the formal frameworks for describing complex problems, we emphasise some of the measurement issues inherent in assessing CPS skills with single tasks (i.e., fixed item difficulty, low or unknown reliability, and a large impact of random errors). With examples of the *MicroDYN* test and the *MicroFIN* test (two instances of the MCS approach), we show how to adequately score problem-solving skills by using multiple tasks. We discuss implications for problem-solving research and the assessment of CPS skills in general.

Keywords: Complex problem solving; Knowledge acquisition; Knowledge application; MicroDYN; MicroFIN; Multiple complex systems

When cognitive scientists want to know how a person copes with certain problems, they cannot just read the person's mind, but rather, they usually have to present the person with a set of valid tasks and assess the problem-solving strategies that he or she applies. In the pioneer era of research on human problem solving, there was a lot of research on rather simple and academic problems such as the *Tower of Hanoi* (Simon, 1975), Duncker's

Correspondence should be addressed to Samuel Greiff, ECCS unit, University of Luxembourg, 6, rue Richard Coudenhove Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg. Phone: +352-466644-9245. Email: samuel.greiff@uni.lu

We are grateful to the TBA group at DIPF (http://tba.dipf.de) for providing the item authoring tool CBA-Item Builder and technical support.

No potential conflict of interest was reported by the authors.

This research was funded by grants from the German Federal Ministry of Education and Research [grant number LSA004], [grant number 01JG1062] and the Fonds National de la Recherche Luxembourg (ATTRACT "ASKI21").

(1945) *Candle Problem*, or the problem of *Missionaries and Cannibals* (Jeffries, Polson, Razran, & Atwood, 1977). The simulation of both realistic and complex problems provided a great step forward in research on human problem solving: With the advent of computers in psychological laboratories during the 1970s, computer simulations of complex scenarios such as *Lohhausen* (Dörner, Kreuzig, Reither, & Stäudel, 1983), *Milk Truck* (Schunn & Klahr, 2000), or the *Sugar Factory* (Berry & Broadbent, 1984) became increasingly popular in the scientific community as methods for examining human problem solving and decision making in realistic tasks (i.e., microworlds; Papert, 1980, p. 204) while still having the advantage of standardised laboratory conditions. For instance, Dörner (1989) elaborated on systematic human failures in coping with complexity, whereas Berry and Broadbent (1984) did research on the influence of implicit knowledge on complex system control, and Klahr and Dunbar (1988) focused on scientific discoveries and hypothesis testing in complex environments.

Complex problems (or microworlds; Kluge, 2008) seem to have greater ecological validity than other cognitive tasks such as tasks used in classical tests of intelligence (Beckmann, 1994). In complex microworlds, problem solvers can manipulate certain *input variables* and observe the resulting changes in a set of *outcome variables*. While doing so, problem solvers have to *acquire* and *apply* knowledge about the complex scenario's structure in order to reach their goals (i.e., build a *representation* of the problem and search for a *solution*; Novick & Bassok, 2005), and this involves processes such as information reduction (Klauer, 1993), causal learning via interaction (Bühner & Cheng, 2005), hypothesis testing (Klahr & Dunbar, 1988), dynamic decision making (Edwards, 1962), and self- and task-monitoring (Osman, 2010).

But even if these processes of knowledge acquisition and knowledge application seem to be highly relevant for problem solving in various domains of daily life such as academic (e.g., Wüstenberg, Greiff, & Funke, 2012) or occupational success (Danner, Hagemann, Schenkin, Hager, & Funke, 2011), research on complex microworlds has faced some major issues that could not be sufficiently solved until now: There was (1) a *lack of comparability between different microworlds*: In early research on complex problem solving (CPS), different opinions about how to define "complexity" (Quesada, Kintsch, & Gomez, 2005) as well as a variety of different scenarios such as Lohhausen, Milk Truck, and the Sugar Factory emerged, and it became difficult to determine the common attributes of those complex problems and to compare them with each other directly (Funke, 2001). Adding to this, (2) *scalability remained unclear* as single time-consuming simulations (e.g., the time-on-task for Lohhausen was about 16 hr; see Dörner et al., 1983, p. 120) were used to measure CPS skills, and different measures of performance in different microworlds did not necessarily correlate with each other or with traditional measures of general mental ability (Dörner, 1986;

Wenke, Frensch, & Funke, 2005) even if there was considerable conceptual overlap between performance on CPS tasks[1] and intelligence tests. So it has previously been unclear whether performance scores across a number of complex problems can be summed to form consistent and homogenous scales.

One decade ago, Funke (2001) proposed using formal frameworks to compare different scenarios with respect to the formal features of their causal structures. This solved the first problem (i.e., lack of comparability) but not the issue of scalability. In this paper, we will extend this approach. First, we will briefly provide background information on (1) the concept of domain-general skills, which are relevant for CPS (Fischer, Greiff, & Funke, 2012), and (2) how to design tasks that address these problem-solving skills in such a way that they are comparable with regard to their underlying formal structure (Funke, 2001). We will then (3) outline the most important measurement issues that have resulted from unclear scalability and that have yet to be resolved, and (4) introduce the multiple complex systems (MCS) approach, which is based on formal frameworks, as a viable way to both overcome these measurement issues and enable solid research on problem-solving skills.

THE PROCESS OF COMPLEX PROBLEM SOLVING

According to Mayer and Wittrock (2006), problem solving takes place when a given state has to be transformed into a goal state and no obvious or routine method of solution is available. A problem is *complex* if a sizable number of interrelated factors have to be considered in order to derive a solution (Weaver, 1948). As prior knowledge about complex problems is often false or at least incomplete (Dörner, 1989), the complex problem solver usually attempts (1) knowledge acquisition and (2) knowledge application (cf. Fischer et al., 2012; Funke, 2001; Novick & Bassok, 2005) in order to adequately represent and solve the complex problem in a viable way.

Knowledge acquisition

When confronted with a complex problem, a problem solver has to build a parsimonious and viable representation of the most relevant aspects of the problem structure. That is, he or she first has to acquire viable knowledge about the problem. On the basis of knowledge about (1) possible states of the specific problem at hand, (2) analogous problem structures, or (3) abstract solution schemas (e.g., "vary one thing at a time"; Tschirgi, 1980), an initial assumption about the relevant aspects of the problem and

[1] Throughout this paper, the term "simulation" describes the whole measure of CPS. Different complex systems within a simulation are called "tasks", each of which may contain different "items" to measure different processes such as knowledge acquisition or knowledge application.

hypotheses about how these aspects are interrelated need to be mentally represented (each kind of data representation highlights certain features and distinctions and downplays irrelevant features and distinctions; cf. Newell & Simon, 1972; Schunn & Klahr, 1995). As these initial assumptions are often false or at least incomplete in complex situations (Dörner, 1989), their viability has to be tested by directly interacting with the problem (cf. Klahr & Dunbar, 1988). Each interaction with the system can be seen as an experiment (varying the state of the problem), which generates information that in turn may allow the problem solver to accept, reconsider, or reject the current assumptions (Klahr & Dunbar, 1988). Bühner and Cheng (2005) emphasised the special importance of active interventions for causal learning (in contrast to the mere observation of covariation). The result of effective interactions and learning is a viable mental representation of the most important aspects of the problem's causal structure (i.e., subject-matter knowledge; cf. Even, 1993). Schunn and Klahr (1995) described this process of acquiring subject-matter knowledge as a search through possible experiments, hypotheses, data representations, and experimental paradigms.

Knowledge application

After a sufficient amount of subject-matter knowledge (Even, 1993) has been acquired, a feasible solution has to be derived. Systematically searching for a solution usually implies applying knowledge about (1) prior encounters with similar situations that were successfully solved (cf. instance-based learning theory; Broadbent, Fitzgerald, & Broadbent, 1986; Gonzalez, Lerch, & Lebiere, 2003), (2) the current schematic representation of the problem (Sweller, 1988), or (3) general solution heuristics applicable in the current situation (Gigerenzer & Brighton, 2009; Kahneman, 2011). The specific knowledge, applied in a certain way to structure or constrain the search process, depends on a variety of personal and situational features such as expertise and meta-strategic knowledge (Kuhn, 2000) or the salience of important features (Novick & Bassok, 2005). When a decision to implement an intervention (or a series of interventions) has been made, the solution has to be implemented. At the same time, the consequences of each intervention and the system's autonomous developments have to be monitored as they may have implications for the representation of the system and for future decisions (cf. dynamic decision making; Edwards, 1962).

If the problem solver is unable to find a solution, he or she may switch back to knowledge acquisition: For instance, when the rate of progress is perceived to be too slow to solve the problem in time (MacGregor, Ormerod, & Chronicle, 2001), or when the problem solver gets stuck in an impasse (Ohlsson, 1992), there are often changes in the representation of the problem (e.g., relaxation of constraints) or in the use of strategy (Fischer et al., 2012).

CPS and related constructs

The theoretical distinction between CPS and related constructs such as reasoning, working memory capacity (WMC), or domain-specific problem solving has been investigated frequently (Wittmann & Süß, 1999). Whereas some researchers have highlighted commonalities between the constructs, others have focused on differences. Both reasoning and working memory overlap theoretically with CPS (Bühner, Kröner, & Ziegler, 2008; Kröner, Plass, & Leutner, 2005), but there are also substantial and important conceptual differences (e.g., Schweizer, Wüstenberg, & Greiff, 2013; Wüstenberg et al., 2012).

Reasoning can be broadly defined as the process of drawing conclusions in order to achieve goals, thus informing problem-solving and decision-making behaviour (Leighton & Sternberg, 2004). It has been linked to executive control processes that allow a person to analyse simple problems, create solution strategies, monitor performance, and adapt behaviour accordingly. Interestingly, the skills necessary for CPS are often identified with the same labels as those for reasoning. As outlined above, CPS also requires the acquisition and application of knowledge and the monitoring of behaviour (Funke, 2001), and problem solving is part of almost every definition of reasoning (Sternberg & Berg, 1992). Nonetheless, Raven (2000) separates CPS from reasoning, focusing on the dynamic interactions necessary in CPS for revealing and incorporating previously unknown information as well as for achieving a goal using subsequent steps that depend on previous steps. The major difference between reasoning and CPS is therefore whether or not there is a need for "experimental interaction with the environment" (Raven, 2000). On this basis, CPS and reasoning can be viewed as related but distinguishable constructs. Empirically, this assumption is supported by studies that have reported moderately high correlations between CPS and reasoning (e.g., Danner et al., 2011; Greiff, Wüstenberg et al., 2013; Wüstenberg et al., 2012).

WMC, on the other hand, is defined as the capacity of the cognitive system to simultaneously store and process information (Baddeley, 1989). It is very closely related to reasoning (e.g., Bühner, Krumm, & Pick, 2005; Kyllonen & Christal, 1990) and is a well-established predictor of different higher order cognitive tasks such as language comprehension (Daneman & Merikle, 1996). Concurrently, WMC may limit CPS performance (Bühner et al., 2008; Süß, 1999). To this end, Wirth and Klieme (2003) argue that "in most dynamic problem situations, [. . .] more than one goal has to be pursued. The underlying structure of the problem is complex, and the amount of relevant information exceeds the capacity of working memory" (p. 332). According to this theoretical view, WMC should predict CPS as it limits the amount of information that can be searched, acquired, and applied when solving a complex problem. However, there are clear theoretical differences between

WMC and CPS. Funke (2010) emphasises that CPS cannot be reduced to simple cognitive operations such as a mere sequence of memory processes. Rather, it is to be understood as an organised interplay of simple cognition and different complex cognitive processes, including the self-guided planning, execution, and evaluation of actions and the application of strategies that are implemented to reach one or more overarching goals (Funke, 2010). Accordingly, WMC may be relevant for CPS, but it does not represent a genuine aspect of it. Empirically, the distinction between WMC and CPS is supported by the incremental validity of CPS scores over WMC in predicting school grades (Schweizer et al., 2013) and moderately high correlations between WMC and CPS (Bühner et al., 2008).

Finally, a large amount of research has been conducted on human problem solving and expertise in specific domains, usually referred to as domain-specific problem solving (Sugrue, 1995), including mathematical (e.g., Daniel & Embretson, 2010), scientific (e.g., Dunbar & Fugelsang, 2005), or technical (e.g., Baumert, Evans, & Geiser, 1998) problems. Domain-specific problems, as encountered outside the laboratory, are always embedded semantically, and the success of a problem solver depends on his or her experience and subject-matter knowledge in this specific area (cf. Sugrue, 1995). But of course, there are domain-general mental processes involved in solving problems regardless of the domain. Knowledge acquisition (i.e., building a mental representation) and knowledge application (i.e., finding a solution) are defining components of problem-solving theories in any domain (cf. Mayer & Wittrock, 2006; Novick & Bassok, 2005; Mayer, Larkin, & Kadane., 1984). Funke (2010) argues that complex and general mental processes are highly relevant when solving new problems and, according to him, the use of general mental representation formats such as causal networks are relevant for knowledge acquisition but not bound to specific domains. To this end, Novick, Hurley, and Francis (1999) state that domain-general processes in problem solving are crucial for problem representation because abstract schemas are more useful than specifically relevant example problems for understanding the structure of novel problems. These general representations are not contaminated by specific content and can thus be generalised more easily (Holyoak, 1985). This line of research does not question that domain-specific processes exhibit high relevance in real-life problem solving (e.g., Wason & Shapiro, 1971), but there is still a substantial degree of domain-generality in CPS (Buchner, 1995; Sternberg, 1995). Empirically, Scherer and Tiemann (2012) were able to distinguish domain-specific and domain-general problem solving as related but separate factors.

In summary, basic cognitive abilities such as reasoning and WMC cannot completely account for performance in CPS, and domain-specific knowledge is not sufficient for (but may result from) solving unknown complex problems in any domain. This is in line with the theory of cognitive cascades

(Bornstein, Hahn, & Heynes, 2010; Fry & Hale, 1996), which posits that basic cognitive abilities predict more complex ones. After this initial theoretical classification, the following section will provide an overview of existing measures of domain-general CPS skills.

FORMAL FRAMEWORKS FOR DESCRIBING GENERAL ASPECTS OF COMPLEX PROBLEMS

Over the last 30 years, experimental research has produced a variety of findings on CPS largely by using measures composed of a large number of elements, time-delayed effects, nonlinear relations, and complex structures (e.g., Dörner, 1989). These tasks were often constructed unsystematically and ad hoc. From a psychometric perspective, these measures were prohibitive (Funke, 2001) as they varied considerably with regard to the systems underlying them and their cognitive demands, thus rendering it impossible to compare empirical results across different studies.

In response to these issues in the measurement of domain-general problem-solving skills such as systematic knowledge acquisition and knowledge application, Funke (2001) introduced the formal frameworks of linear structural equations (LSEs) and finite state automata (FSA), which allowed complex problems to be described systematically on a formal level. A coherent formal description of different complex problems ensures a minimal set of commonalities between these problems (instead of mixing apples and oranges) and allows for the systematic comparison of different complex problems with regard to their underlying structure (instead of or in addition to their surface features or semantic context). As these frameworks are an important prerequisite for the CPS measurement approach based on MCS proposed in this paper, we will elaborate further on both frameworks.

Linear structural equations

LSEs describe a framework for modelling linear relations between quantitative variables, such as the influence of coffee consumption on thirst and alertness (within certain boundaries). On a formal level, LSE systems contain a set of input variables (which can be set by the problem solver) and a set of output variables (whose values may linearly depend on other output or input variables) as well as linear relations between these variables. In dynamic systems, an output variable may also influence itself, called eigendynamic (Funke, 2001). In order to solve this kind of problem (e.g., to obtain a certain level of alertness by drinking coffee), a problem solver has to (1) explore the linear relations between input and output variables in a first phase (knowledge acquisition) and (2) reach certain goal values for the output variables in a second phase (knowledge application).

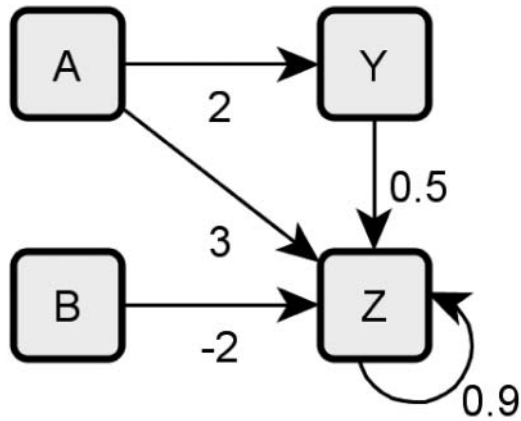

Figure 1. Structure of a linear system (Funke, 2001) with two input variables (A and B), two output variables (Y and Z), and the relations between them (arrows).

Figure 1 illustrates the following system of interrelated variables:

$$Y_{t+1} = 2 \times A_t;$$
$$Z_{t+1} = 0.9 \times Z_t + 0.5 \times Y_t + 3 \times A_t - 2 \times B_t;$$

In this example, varying the exogenous variable A at any discrete point in time t has multiple direct effects on Y and Z and an additional delayed and indirect effect on Z (A influences Y, and Y in turn influences Z). The endogenous variable Z is directly dependent on both itself (eigendynamic) and on three other variables (A, B, and Y).

Of course, less abstract problems can be formulated as LSE systems as well: For example, the complex problem of managing a sugar factory was described as an LSE containing one output variable (sugar production) that was negatively related to itself and positively related to one input variable (number of workers) by Berry and Broadbent (1984). In this system, the problem solver had to acquire knowledge about the causal interrelations of variables and to apply this knowledge in order to successfully control the amount of sugar produced by manipulating the number of workers.

Finite state automata

In contrast to LSE systems, FSA systems are useful for describing relations between qualitative variables, for example, the discrete state changes triggered by the buttons of a mobile phone or a ticket vending machine (Buchner & Funke, 1993). An FSA contains a limited number of states S (e.g.,

TABLE 1
State transition matrix of the system shown in Figure 2 (Funke, 2001)

	Input X to current state S	
Resulting state S	X1	X2
S0	S1	S0
S1	S2	S2
S2	S0	S2

"on" and "off") and a limited number of interventions X (e.g., buttons) as well as a function that specifies the state following each possible other state and/or intervention (see Table 1). Whereas the distinction between the two frameworks is pragmatic to a large degree as the two representations can be translated into each other (e.g., Cohen, 1968), it is important to note that an FSA system differs from an LSE system where the states change quantitatively and discretely and are therefore not limited to certain qualitative categories. LSEs can therefore be considered a special form of FSA with a very large number of ordinal categories, which would be highly impractical to represent in FSA (Neubert, Kretzschmar, Wüstenberg, & Greiff, 2014). The problem solver is not shown the states of an automaton directly, but there is a visual output (e.g., on the screen of a mobile phone) based on the current state or the current state transition of the automaton (Funke, 2001). In order to control an unknown FSA, the problem solver has to (1) acquire knowledge about the consequences of interventions as well as their conditions and (2) apply this knowledge in a goal-oriented way to reach a certain state.

For instance, Table 1 illustrates a simple finite automaton with two buttons (X1, X2) that cause a state transition to one of three different states (S0, S1, S2) depending on the current state of the automaton (this representation is called the state transition matrix; Funke, 2001). In Figure 2, the same automaton is visualised as a network diagram containing three nodes

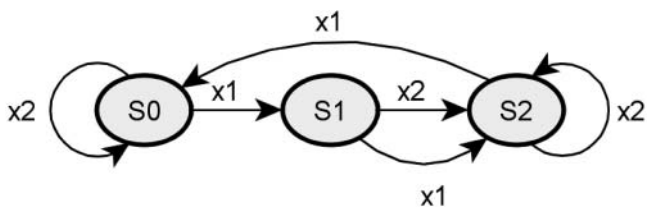

Figure 2. Graphical representation of a finite state automaton (Funke, 2001) with three states (Z1, Z2, Z3) and two possible inputs (X1, X2) that lead to state transitions (arrows).

(S0, S1, S2) with two arrows (X1, X2) pointing from each node to the next node.

Nearly every problem can be approximated by a set of possible states and state transitions. Thus, the FSA framework may be especially suited for problem-solving research (for more detail on this concept, see the FSA *Space Shuttle*, which was used in the Programme for International Student Assessment; Wirth & Klieme, 2003).

In general, the two formal frameworks introduced by Funke (2001), LSE and FSA systems, solved the lack of *comparability* in CPS research by specifying commonalities and discrepancies between all complex problems that can be formulated within a common framework. Existing microworlds could now be compared, and new microworlds could be designed with regard to the underlying causal structure of a problem. Therefore, the formal frameworks are widely used in problem-solving research (e.g., Funke, 2001; Kluge, 2008; Kröner et al., 2005; Wüstenberg, et al., 2012) and provide an important prerequisite for the MCS approach, which addresses the issue of the unclear *scalability* of complex problems and which we will present in this paper.

MEASUREMENT ISSUES IN CPS

The formal frameworks introduced by Funke (2001) allowed different complex scenarios to be compared on the basis of their underlying structure (e.g., Greiff & Funke, 2010) and not just on the basis of surface features and fuzzy descriptions of problems. But even though Funke's (2001) approach solved the lack of comparability, the *scalability* of CPS skills remained unknown because all complex problems available at the time shared a single major shortcoming: They consisted of only a single task or problem (e.g., *Tailorshop*; Funke, 2003). More specifically, single-task testing causes certain characteristic problems:

(1) There is no *variation in difficulty* across tasks as only a single task is used (in fact, difficulty is often not even reported; Greiff, Wüstenberg, & Funke, 2012). That is, system structure and other task characteristics remain constant, which results in different discrimination between groups of low, average, and high performers. For instance, an item with average difficulty in a sample of low performers may be too easy for a sample of high performers and unable to discriminate different levels of ability.
(2) Single-task testing results in *low and even unknown estimates of reliability*. Reliability can be estimated adequately only when there are multiple tasks that can be assumed to measure the same construct (or the same task multiple times). The few studies that have

conducted reliability estimates for single CPS tasks using retests have reported considerably low estimates with r_{tt} values ranging from .56 to .69 (Süß, 1996). As the square root of reliability marks the upper bound of validity, and reliability increases with the number of tasks (Carmines & Zeller, 1991), single-task testing may underestimate the validity of a test due to small or unknown reliabilities.

(3) One single random error—in particular, at the beginning of a CPS task—can *heavily compound performance* and lead to low test scores even when ability on the underlying construct is high. For instance, in the CPS task *Tailorshop* (Funke, 2003), one substantial mistake in the beginning (e.g., a random typing error) irreversibly affects all subsequent steps as well as the final outcome. The same mistake at the end of the test may impact performance less or differently.

As these measurement issues preclude a meaningful interpretation of many empirical findings on the topic, Süß (2001) concluded that the importance of the theoretical construct CPS has been difficult to evaluate until now. The construct validity of many operationalisations of CPS is difficult to estimate, especially due to unknown and low reliabilities. In this paper, we want to contribute a solution to these problems: We will present *multiple-task testing*, which is based on the two formal frameworks proposed by Funke (2001) as a means for overcoming (1) unvaried difficulty, (2) low and unknown reliability, and (3) the overweighting of random errors—the three major issues resulting from single-task testing. We will now introduce and discuss the MCS approach within formal frameworks as an approach for building multiple-item scales that can be used to overcome the measurement issues mentioned above.

THE ADVENT OF MULTIPLE COMPLEX SYSTEMS

MCS are based on formal frameworks but extend both LSEs and FSA by including measurement considerations that have the potential to solve the characteristic problems of single-task testing mentioned above.

Important first steps toward tackling the problems of single-task testing were already made with the introduction of microworlds such as the finite state automaton *Space Shuttle* (Wirth & Klieme, 2003) or *Multiflux* (Kröner et al., 2005). Unlike classical measures of CPS in which trial-and-error behaviour at the beginning of the task influenced the final problem-solving score, these more recent CPS tasks included an evaluation-free exploration phase. This solved the problem of initial random behaviour influencing the final problem-solving scores by separating the processes of knowledge acquisition and knowledge application. As real-world problem solving allows for repeated alternation between the two processes, this separation between

knowledge acquisition and knowledge application is a compromise between the psychometric assessment of CPS and ecological validity (Kröner et al., 2005). Separating the two processes allows for their independent assessment even though it may be at odds with real-life problem solving.

In addition, providing feedback about the correct solution after the knowledge acquisition phase solved the problem of the knowledge application scores being confounded with individual differences in knowledge acquisition. This allowed for the distinct measurement of knowledge acquisition and knowledge application. Multiple items were used to ask participants about their knowledge of the underlying structure of the system as well as their ability to control it. For example, after exploring the *Space Shuttle* for 20 min, problem solvers had to answer approximately 20 items about the underlying logic as an assessment of their knowledge acquisition. Similarly, multiple items requiring participants to direct the system toward producing certain values were used to assess knowledge application. The *Multiflux* simulation is very similar in that, after an initial exploration phase, several items are presented to assess knowledge acquisition and knowledge application. This simulation also provides participants with the correct structural diagram underlying the simulation after the knowledge acquisition phase. In summary, these innovations in CPS measurement solved the problem of initial random behaviour influencing the final problem-solving scores and allowed for the distinct measurement of knowledge acquisition and knowledge application.

Whereas this was certainly an improvement over microworlds that did not include an evaluation-free exploration phase and that did not make a clear distinction between knowledge acquisition and knowledge application (Kröner et al., 2005), it is also important to note that all items in these microworlds were based on the very same underlying task structure. Obviously, dependencies between these items could arise as a problem solver might understand the system as a whole and would then be more likely to answer all items correctly (Greiff et al., 2012).

The number of tasks within these measures of CPS, however, is limited by the assumption that microworlds need substantial time spent on a task to sufficiently model reality (ranging from at least 30 min up to several days; e.g., Frensch & Funke, 1995). Consequently, microworlds such as *ColorSim* (Kluge, 2008), *Space Shuttle* (Wirth & Klieme, 2003), or *Multiflux* (Kröner et al., 2005) require a minimum of 30 min of processing time, which, from a practical perspective, limits the number of employable tasks to one (Greiff et al., 2012). In MCS processing, the required time is reduced, and thus a sufficient number of less time-consuming and independent tasks can be presented. That is, in line with simulations such as *Multiflux*, the MCS approach improves upon classical measures by including an evaluation-free exploration phase and feedback on the correct solution after the knowledge

acquisition phase. Furthermore, the MCS approach improves upon other recent measures of CPS by employing an entire set of several independent tasks and allowing the researcher to (1) vary the difficulty of both knowledge acquisition and knowledge application, (2) estimate and increase reliability, and (3) lessen the impact of single random errors.

On the basis of the formal frameworks proposed by Funke (2001), we designed MCS, each solvable within a short amount of time. The underlying task structures in MCS can be described by either LSEs or FSA. The first MCS approach based on LSEs is known as *MicroDYN* and the second MCS approach based on FSA is called *MicroFIN*. "Micro" in both cases refers to the shorter time on task and the limited number of elements in a task's structure, whereas DYN alludes to *DYNAMIS*, the name given to the first LSE approach by Funke (2001), and FIN alludes to finite state automata as the underlying formalism (Greiff, Fischer, et al., 2013).

Multiple complex systems in LSEs: *MicroDYN*

As MCS formulated as LSEs are comparable by definition, we created a measure composed of multiple CPS tasks that reflected the defining theoretical aspects of CPS: (1) The acquisition of knowledge about how to adequately represent the problem and (2) the application of this knowledge to solve the problem. Consequently, within *MicroDYN*, problem solvers are instructed to perform two subtasks (items), each of which is addressed in a separate phase with 5 min of time-on-task overall: In Phase 1, knowledge acquisition (3 min), respondents explore the task and represent their acquired knowledge by manipulating inputs and deriving conclusions from their individual manipulations. In Phase 2, knowledge application (2 min), respondents have to achieve predefined target values in the output variables by correctly manipulating the input variables within a limited number of active interventions. Usually, in about 1 hr of testing time, a set of approximately 10 *MicroDYN* scenarios is administered, yielding 10 independent measurement points (in comparison with only one in single-task testing). As an illustration, a screenshot of a typical *MicroDYN* task, *handball team*, with three input and three output variables is depicted in Figure 3. There, different kinds of training labelled Training A, B, and C serve as input variables, whereas different team characteristics labelled Motivation, Power of the Throw, and Exhaustion serve as output variables.

As research on CPS is particularly focused on domain-general cognitive processes (see section above; cf. Kröner et al., 2005; Raven, 2000), semantic cover stories in *MicroDYN* activate as little subject-matter knowledge as possible and are varied between items (Greiff et al., 2012). Highly different semantic covers are used in each *MicroDYN* task; for example, coaching a sports team (see the handball training task in Figure 3), feeding an alien

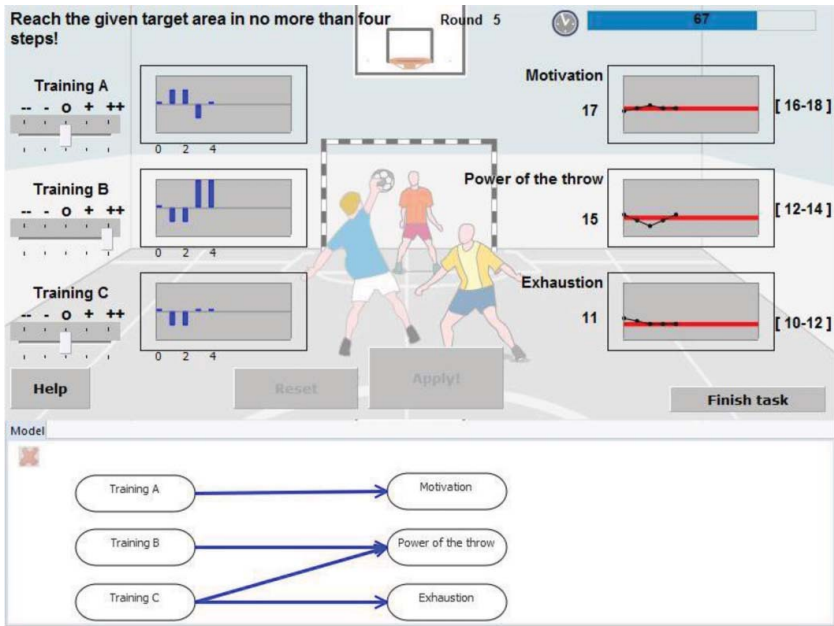

Figure 3. Screenshot of the knowledge application phase within a *MicroDYN* task. The left side of the screen depicts sliders for manipulating input variables (Training A, Training B, Training C), and the right side depicts current and goal values for output variables (Motivation, Power of the Throw, Exhaustion). The correct causal model is at the bottom of the screen (cf. Wüstenberg et al., 2012).

creature, or driving a moped. To prevent uncontrolled influences of subject-matter knowledge, input and output variables are labelled either without deep semantic meaning (e.g., Training A) or fictitiously (e.g., Wildvine as the name of a flower or Natromic for a fertiliser). Cover stories are realistic and semantically rich, but they do not provide information about how to solve the specific problem at hand nor do they activate helpful subject-matter knowledge.

An additional asset in *MicroDYN* is the ability to scale the difficulty, which is related to variations in task characteristics: Whenever a new task out of the set of independent *MicroDYN* tasks is administered, difficulty can be decreased or increased by varying the underlying system structure. Greiff and Funke (2010), Greiff, Krkovic, and Nagy (2014), as well as Kluge (2008) provided the first empirical insights into which task characteristics (e.g., degree of connectivity, direct and indirect effects) need to be varied to systematically and predictably change the task difficulty.

In the second problem-solving phase, knowledge application, system interventions are targeted toward reaching a specific goal state in the LSE

(e.g., a high level of Motivation or a low level of Exhaustion in Figure 3). This allows for a direct evaluation of whether respondents have reached the goal or not, whereas a number of options exist for how to check for the correct representation of a problem in the first phase of knowledge acquisition. To this end, Funke (2001) introduced causal models in which participants are instructed to draw lines between variables indicating the amount of knowledge that the participants generated. These models can then be compared with the correct causal models of the underlying task structure (see Figure 3). However, the scientific community has proposed other forms of assessment including multiple-choice questions about the structure of the problem (e.g., Kluge, 2008; Kröner et al., 2005) or constructed responses (Frensch & Funke, 1995). The issue of how problem solvers' performance is reflected in overt behaviour is closely related to the question of how to transform data generated by problem solvers (e.g., the causal model drawn, the distance between given and achieved goals) into specific indicators and scores. Options include, for instance, continuous indicators of problem representation in the first phase in which different types of mistakes are compared and weighted differently (Funke, 2001); indicators rooted in signal detection theory, combining misses, false alarms, hits, and correct rejections into sensitivity and bias scores (Beckmann, 1994); logarithmic deviation scores between given and achieved goal states (Kluge, 2008); or categorical scoring schemes (Wüstenberg et al., 2012) for solution patterns in the second phase. The notable difference between classical single-task testing and MCS in *MicroDYN*, however, is that only one (independent) performance indicator for each of the two phases is available in single-task testing, and this can easily be impaired by external disturbances; whereas in *MicroDYN*, each task yields two indicators, summing to approximately 10 knowledge acquisition scores and 10 knowledge application scores, depending on the specific number of tasks employed in a set of *MicroDYN* tasks.

Multiple complex systems in FSA: *MicroFIN*

In *MicroFIN* (Greiff, Fischer, et al. 2013; Neubert et al., 2014), the MCS principle of administering an entire set of tasks with a short processing time and a reduced number of elements is applied to the formal framework of FSA. Comparable to *MicroDYN*, a first phase in which respondents are instructed to freely explore the complex system and to provide data on the knowledge they acquire during this process, is followed by a second phase, in which respondents apply their knowledge to reach predefined goal states. Testing time (approximately 1 hr) and number of tasks in a *MicroFIN* set (approximately 10) are comparable to *MicroDYN*, extending the approach of MCS not only to LSEs, but also to FSA.

Figure 4. Screenshot of the knowledge acquisition phase within a *MicroFIN* task. Possible inputs are one of three settings for different types of laundry (A, B, C), the position of three different slides (red, yellow, blue), or a click of the "start" button.

Figure 4 illustrates the principle of *MicroFIN*. There, the task "washing machine" is displayed during the second phase of knowledge application. In this automaton, the test taker must find out about an unknown technical device that is rather complex because the desired goal state (clean laundry) depends on the interaction of different settings. Whereas in *MicroDYN* (Figure 3), elements are related to each other in a quantitative way, relations between states in *MicroFIN* are of a qualitative nature, and this constitutes the main difference between the LSE and FSA frameworks (Neubert et al., 2014). In a set of *MicroFIN* tasks, semantic covers can be varied and designed to activate as little subject-matter knowledge as possible (i.e., inputs are not labelled in a meaningful manner; their effects have to be explored in the knowledge acquisition phase) while simultaneously simulating a motivating and realistic problem. The underlying states and transitions are changed in order to vary the difficulty levels even though little is known about how specific task characteristics impact difficulty. However, the complete formal description of *MicroFIN* tasks within the FSA framework provides the background necessary for systematically varying task difficulty (Buchner, 1995; Funke, 2001; Neubert et al., 2014).

As in *MicroDYN*, *MicroFIN* enables a range of possibilities for recording problem solvers' performance and for transforming performance data into specific indicators. Whereas the second theoretical process, the search for a solution, is measured in a straightforward way in *MicroFIN* by setting a specific goal state and instructing respondents to move toward it, a variety

of options have been suggested for measuring the first process, knowledge acquisition, in FSA. For instance, in a manner that is equivalent to causal models in LSEs, Buchner (1995) suggested that individual transition matrices be assessed as a way to reflect knowledge about a problem's representation and that such matrices then be compared with the actual transition matrices. That is, the smaller the difference between an individual and actual matrix, the better and the more complete the knowledge a respondent has gathered. Further, either as constructed responses or multiple-choice questions, different types of judgement tasks (predictive, interpolative, and retrognostic; Buchner, 1995) or verification tasks (Buchner & Funke, 1993) can be used to measure knowledge acquisition. The application of optimal solution sequences and the distance to a specific goal state in terms of the number of missing steps until the goal would have been reached are well-established indicators of knowledge application (Buchner, 1999). In line with the MCS approach, each of the approximately 10 *MicroFIN* tasks in a complete set yields two indicators, one on knowledge acquisition and one on knowledge application.

Both *MicroDYN* and *MicroFIN* are aimed at advancing LSEs and FSA as well-established formal frameworks in problem-solving research (Funke, 2001) designed to measure knowledge acquisition and knowledge application in CPS even though the two formalisms differ substantially. For instance, the representation of knowledge in *MicroFIN* is essentially different from *MicroDYN* as effects of inputs in *MicroFIN* always depend on inner states of the task while they are assumed to be the main effects in *MicroDYN*. Despite these differences, substantial empirical correlations between LSEs and FSA show that the two formalisms tap into the same underlying construct (Greiff et al., 2012; Greiff, Fischer et al., 2013). We will now describe what specific advantages are to be expected when extending LSEs and FSA toward *MicroDYN* and *MicroFIN* within the MCS approach versus single-task testing.

Advantages of multiple complex systems

Using MCS avoids single-task testing per definition, and thereby *MicroDYN* and *MicroFIN* finally provide solutions for the characteristic weaknesses of single-task tests mentioned above (i.e., the lack of variation in difficulty, low or unknown reliability, and a large influence of random errors; cf. section measurement issues in CPS):

(1) In MCS, there can be *variation in item difficulty*: As every complex system can have a different difficulty, the problem-solving performance of low, average, and high performers can be examined with adequate discrimination within the MCS approach. Whereas prior research that was based on single-task testing often applied single

very difficult tasks and thereby focused on general human problems in coping with a single difficult problem (e.g., Dörner, 1989), the MCS approach provides a broader picture. For instance, Greiff et al. (2012) reported that *MicroDYN*-task item difficulties varied between $p = .04$ and $p = .69$ for knowledge acquisition and knowledge application, which implies that a person with average skills will probably fail to solve some but not all of the problems and will be able to solve the other ones (which is much more informative than knowing that he or she was unable to solve a single difficult item). By implementing multiple items with varying difficulties (see Figure 5 for a set of *MicroDYN* MCS that varied with regard to both the number and interconnectedness of elements), we can assess problem-solving skills on different levels of performance. It is also possible to systematically examine (and control for) effects of item difficulty on the relations of CPS to other constructs (e.g., Kluge, 2008) or effects of problem features on item difficulty (e.g., Greiff & Funke, 2010).

(2) In MCS, *reliability can be determined and enhanced*: As there are multiple independent items, we can calculate adequate estimates of reliability (e.g., split-half reliability and internal consistencies). By

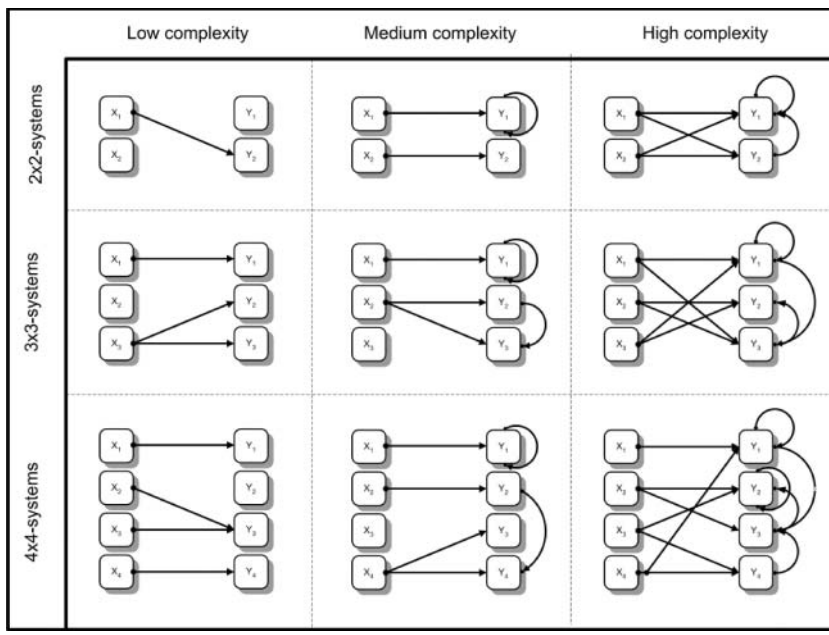

Figure 5. Multiple complex systems with varying numbers of elements (rows) and varying complexity (interconnectedness and direct/indirect effects; columns) (taken from a presentation by Greiff, Wüstenberg, & Funke, 2011).

contrast, estimates of reliability reported for single-task testing have often been inflated because the items depended on the prior solution to a single problem (i.e., correlations between items can depend on this prior solution and not only on a latent ability). Adding to this, reliability can be increased in the MCS approach by adding tasks to a test (Carmines & Zeller, 1991): For a *MicroDYN* set of 11 tasks, Greiff et al. (2012) reported reliability estimates between $\alpha = .85$ and $\alpha = .95$ (similar results were reported by Wüstenberg et al., 2012, for 8 tasks with $\alpha = .73$ to $\alpha = .86$; and Sonnleitner et al., 2012, for 16 tasks with $\alpha = .73$ to $\alpha = .86$). As the square root of reliability marks the upper bound of validity, estimates of validity (e.g., correlations between *MicroDYN* indicators and other constructs such as intelligence) may be adequately corrected for attenuation in the MCS approach when considering relations between CPS and other constructs on a latent level. This addresses a major shortcoming of prior single-task testing, as tests with low and unknown reliability may have resulted in severe underestimations of validity (Süß, 1996). Recent findings on the construct validity of CPS measured by MCS will be reported below.

(3) In MCS, *measurement error is less likely to compound a person's performance*. A severe erroneous decision in a single task does not imply a poor solution in other tasks. However, in single-task testing, such a single decision may automatically result in a low estimate of a person's ability. In the MCS approach, items are independent from each other, and performance on each item does not depend on performance on previous items. Whereas a single random misperception, which is not related to a person's ability per definition, can heavily compound performance in controlling a single finite automaton, this is less likely to occur when controlling multiple automatons in *MicroFIN*. This avoids the overweighting of specific person-item interactions and accommodates the stochastic relation between a person's ability and his or her item response (Rasch, 1980).

(4) Because different tasks impose different strategy requirements during exploration (Wirth, 2004), the use of multiple small tasks in MCS allows for the creation of CPS measures that require *specific strategies* (e.g., Wüstenberg, Stadler, Hautamäki, & Greiff, 2014). For instance, variables affecting themselves (eigendynamics; Figure 1) can be detected only by not manipulating any variables to explore the system's impetus. Thus, the decision to include or exclude eigendynamics in the CPS measure determines whether or not the use of this specific strategy can be assessed. Other tasks may require different specific strategies without which they cannot be solved; an example here is the "vary-one-thing-at-a-time" strategy

(Tschirgi, 1980). As a consequence, MCS tasks can be selected to assure a content-valid strategy assessment to potentially increase their content validity or to obtain information about specific deficiencies in CPS skills; such information may be used for training or developmental purposes.

DISCUSSION: WHERE IS COMPLEX PROBLEM SOLVING HEADED?

In the current paper, we demonstrate how an assessment of domain-general CPS skills is based on computer-based simulations of complex problems. We propose that an adequate assessment of skills (e.g., knowledge acquisition and knowledge application; Funke, 2001) requires a set of multiple problems that are comparable on a formal level (Funke, 2001). For this purpose, we outlined the formal frameworks of LSE systems and FSA (proposed by Buchner & Funke, 1993, and by Funke, 2001). We introduced *MicroDYN* (based on LSEs) and *MicroFIN* (based on FSA) in order to demonstrate how MCS can be applied and scored to overcome some important measurement issues in CPS research. Specifically, the MCS approach provides the following: (1) Different skill levels can be measured with adequate discrimination due to a wide range of task difficulties, (2) reliability and validity can be estimated by including an adequate number of conceptually independent tasks, (3) multiple independent indicators of CPS skills can be included, lessening the impact of single errors during testing, and (4) the specific strategies tailored to specific needs in training or developmental contexts can be implemented and assessed.

Both knowledge acquisition and knowledge application skills seem paramount for solving complex problems and can be reliably addressed within the MCS approach. General mental abilities can be considered an important prerequisite for CPS, but the skills involved in knowledge acquisition (e.g., knowing a systematic strategy of hypothesis testing, deductively generating hypotheses, representing information in a causal network diagram, etc.) and knowledge application (e.g., considering the consequences of one's actions and eigendynamics, adapting plans to recent developments, etc.) provide added value for coping with complex problems (e.g., Wittmann & Hattrup, 2004) and represent a defining part of problem-solving competency (Greiff & Fischer, 2013; Snow, 1989). It seems that, after the construct validity of different indicators of CPS skills has been questioned for a long time (Süß, 2001), we are now able to address the question with adequate methodology.

Still, several urgent questions regarding CPS remain. One line of current CPS research is addressing developmental issues by examining the plasticity of the construct and the influence of lifelong learning and training (e.g., the European Life Long Learning project; www.lllightineurope.com). In

addition, it is necessary to improve our understanding of the specific nature of CPS and its relations to other facets of cognitive performance. Whereas we have some understanding of how CPS is related to general conceptions of reasoning or fluid intelligence (Horn & Cattell, 1966; McGrew, 2009), the role of crystallised intelligence (i.e., specific acquired knowledge; McGrew, 2009) in CPS is still not clear. System-specific knowledge was found to be highly relevant for some classical measures of CPS (Wittmann & Süß, 1999). However, more recent measures of CPS (including MCS) use arbitrary contexts so that the impact of previous knowledge about the system's framework should be substantially reduced (e.g., Greiff, Wüstenberg et al., 2013; Kröner et al., 2005; Wagener, 2001). Abstract knowledge of strategies, such as systematic control of variables (e.g., Wüstenberg et al., 2014) or dynamic systems (e.g., "the robust beauty of linear systems"; Dawes, 1979), on the other hand, might substantially influence problem-solving behaviour. The simple application of known strategies would facilitate the problem-solving process. Thus, further theoretical and empirical work is necessary in order to fully integrate CPS into more thorough frameworks of cognitive abilities such as the Cattell—Horn-Caroll theory (McGrew, 2009).

Other influences might derive from individual differences among participants. For instance, Wittmann and Hattrup (2004) reported performance differences between men and women on the knowledge acquisition dimension of CPS in multiple measures of CPS. The authors thereby speculated that higher risk-aversiveness could cause female participants with the same level of fluid intelligence to implement more cautious interventions (Wittmann & Hattrup, 2004), which may lead to less informative reactions of the system and fewer opportunities to learn about the causal structure of the system (Wüstenberg et al., 2012). As this finding illustrates, the role of interindividual differences (e.g., personality or motivation) is not yet sufficiently understood, and further research on this topic is required (Marshalek, Lohman, & Snow, 1983).

Finally, the mostly numerical (*MicroDYN*) or figural (*MicroFIN*) feedback participants receive may limit the domain-generality of the CPS measures. In line with the construction of the Berlin model of intelligence structure test (Jäger, 1973), in which complex reasoning tasks for the verbal, numerical, and figural content or domain are aggregated to diminish the domain-specific variance in the content and boost the variance of more operation-like abilities, it would be interesting to aggregate the performance scores that are based on different tasks with numerical, figural, or verbal feedback. For instance, it would be possible to present the current states of the outcome variables in *MicroFIN* as a vector of numbers rather than with a graphical representation. Similarly, feedback in *MicroFIN* could be verbal rather than numerical. This would provide important additional support for the notion of the domain-general measurement of CPS.

CONCLUSION

By providing reliable measures of problem-solving skills, the MCS approach is an important step forward in problem-solving research: Since research on problem solving had its zenith in the seventh decade of the last century (e.g., Newell & Simon, 1972), it seems to have faced a major decline in interest in the research community, partly due to methodological issues (Ohlsson, 2012) and a lack of ecological validity (Dörner, 1986). But as the phenomenon of problem solving itself has remained an interesting one to explore (in fact, CPS as a nonroutine behaviour may be increasingly important for today's workplaces; Autor, Levy, & Murnane, 2003), the research community is in need of new forms of tests that can capture the complex processes that occur during problem solving (Rigas & Brehmer, 1999). In recent years, a second wave of interest in problem solving seems to have begun: As is indicated by both the research endeavours reported above and the assessment of problem-solving skills in international large-scale assessments such as the initiative for the assessment of twenty-first-century skills (Griffin, McGaw, & Care, 2011), PISA (2003, 2012, 2015), or the Programme for the International Assessment of Adult Competencies (PIAAC), the assessment of problem-solving skills is increasingly recognised as an important issue in daily life today, even outside of educational research. While the formal frameworks proposed by Funke (2001) have been applied in large-scale assessments since PISA 2003 (see above), the current assessment of CPS skills in PISA 2012 were measured by *MicroDYN* and *MicroFIN* (Organisation for Economic Co-operation and Development [OECD], 2010), that is, within the MCS approach. Based on the formal frameworks proposed by Funke (2001) and the MCS approach outlined in this paper, the assessment of problem-solving skills may be facing its second youth, and only time will tell if it lasts.

Manuscript received 7 December 2012
Revised manuscript received 7 November 2014
Revised manuscript accepted 13 November 2014
First published online 19 December 2014

REFERENCES

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

Baddeley, A. (1989). The uses of working memory. In P. R. Solomon, G. R. Goethals, C. M. Kelley, & B. R. Stephens (Eds.), *Memory: Interdisciplinary approaches* (pp. 107–123). New York, NY: Springer.

Baumert, J., Evans, R. H., & Geiser, H. (1998). Technical problem solving among 10-year-old students as related to science achievement, out-of-school experience, domain-specific control beliefs, and attribution patterns. *Journal of Research in Science Teaching, 35*(9), 987–1013.

Beckmann, J. F. (1994). *Lernen und komplexes Problemlösen* [Learning and complex problem solving]. Bonn: Holos.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology, 36A*, 209–231.

Bornstein, M. H., Hahn, C. S., & Haynes, O. M. (2010). Social competence, externalizing, and internalizing behavioral adjustment from early childhood through early adolescence: Developmental cascades. *Development and Psychopathology, 22*(4), 717–735.

Broadbent, D., Fitzgerald, P., & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77*, 33–50.

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: Lawrence Erlbaum.

Buchner, A. (1999). Komplexes Problemlösen vor dem Hintergrund der Theorie finiter Automaten [Complex problem solving and the theory of finite state automata]. *Psychologische Rundschau, 50*, 206–212.

Buchner, A., & Funke, J. (1993). Finite state automata: Dynamic task environments in problem solving research. *The Quarterly Journal of Experimental Psychology, 46A*, 83–118.

Bühner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 143–168). Cambridge: Cambridge University Press.

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

Bühner, M., Krumm, S., & Pick, M. (2005). Reasoning = working memory≠ attention. *Intelligence, 33*(3), 251–272.

Carmines, E. G., & Zeller, R. A. (1991). *Reliability and viability assessment*. Thousand Oaks, CA: Sage.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*(6), 426.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422–433.

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement, 34*(5), 348–364.

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.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571.

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

Dörner, D. (1989). *Die Logik des Misslingens. Strategisches Denken in komplexen Situationen* [Logic of failure. Strategic thinking in complex situations]. Hamburg: Rowohlt.

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (1983). *Lohhausen: Vom Umgang mit Komplexität* [Lohhausen: On handling complexity]. Bern: Huber.

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. L. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 705–725). New York, NY: Cambridge University Press.

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58*(5), i–113.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4*, 59–73.

Even, R. (1993). Subject-matter knowledge and pedagogical content knowledge: Prospective secondary teachers and the function concept. *Journal of Research in Mathematics Education, 24*(2), 94–116.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving, 4*(1), 19–42.
Frensch, P. A., & Funke, J. (Eds.). (1995). *Complex problem solving. The European perspective*. Hillsdale, NJ: Lawrence Erlbaum.
Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science, 7*(4), 237–241.
Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning, 7*(1), 69–89.
Funke, J. (2003). *Problemlösendes Denken* [Problem solving thinking]. Stuttgart: Kohlhammer.
Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing, 11*(2), 133–142.
Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107–143.
Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*, 591–635.
Greiff, S., & Fischer, A. (2013). Der Nutzen einer Komplexen Problemlösekompetenz: Theoretische Überlegungen und empirische Befunde [Usefulness of complex problem solving competency: Theoretical considerations and empirical results]. *Zeitschrift für Pädagogische Psychologie, 27*(1), 27–39.
Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence, 41*, 579–596.
Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme [Systematic research on complex problem solving by means of minimal complex systems]. *Zeitschrift für Pädagogik, 56*, 216–227.
Greiff, S., Krkovic, K., & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56*(1), 83–103.
Griffin, P., McGaw, B. & Care, E. (2011). *Assessment and teaching 21st century skills*. Heidelberg: Springer.
Greiff, S., Wüstenberg, S., & Funke, J. (2011, April). *Measuring a cross-curricular skill: Competence models and construct validity of dynamic problem solving*. Paper presented at the Conference of the American Educational Research Association in New Orleans, USA.
Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new measurement perspective. *Applied Psychological Measurement, 36*(3), 189–213.
Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapo, B. (2013). Complex problem solving in educational settings – something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*(2), 364–379.
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.
Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253–270.
Jäger, A. O. (1973). *Dimensionen der Intelligenz*. Göttingen: Hogrefe Verlag für Psychologie.
Jeffries, R. P., Polson, P. G., Razran, L., & Atwood, M. E. (1977). A process model for missionaries: Cannibals and other river-crossing problems. *Cognitive Psychology, 9*, 412–440.
Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus and Giroux.
Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48.

Klauer, K. C. (1993). *Belastung und Entlastung beim Problemlösen. Eine Theorie des deklarativen Vereinfachens* [Charge and discharge in problem solving. A theory of declarative simplification]. Göttingen: Hogrefe.

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

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

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

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433.

Leighton, J. P., & Sternberg, R. J. (Eds.). (2004). *The nature of reasoning*. Cambridge: Cambridge University Press.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information-processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 176–201.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*(2), 107–127.

Mayer, R. E., Larkin, J., & Kadane, J.B. (1984). A cognitive analysis of mathematical problem solving ability. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 231–273). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Neubert, J. C., Kretzschmar, A., Wüstenberg, S., & Greiff, S. (2014). Extending the assessment of complex problem solving to finite state automata: Embracing heterogeneity. *European Journal of Psychological Assessment*. Advanced online publication. doi: 10.1027/1015-5759/a000224

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). 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: Cambridge University Press.

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

Organisation for Economic Co-operation and Development. (2010). *PISA 2012 problem solving framework* (draft for field trial). Paris: Author.

Ohlsson, S. (1992). Information processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking*. London: Harvester Wheatsheaf.

Ohlsson, S. (2012). The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. *Journal of Problem Solving, 5*(1), 101–128.

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin, 136*, 65–86.

Papert, S. (1980). Computer-based microworlds as incubators for powerful ideas. In R. Taylor (Ed.), *The computer in the school: Tutor, tool, tutee* (pp. 203–210). New York, NY: Teacher's College Press.

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

Quesada, J., Kintsch, W., & Gomez, E. (2005). Complex problem solving: A field in search of a definition? *Theoretical Issues in Ergonomic Science, 6*(1), 5–33.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Rigas, G., & Brehmer, B. (1999). Mental processes in intelligence tests and dynamic decision making tasks. In P. Juslin & H. Montgomery (Eds.), *Judgement and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 45–65). Hillsdale, NJ: Lawrence Erlbaum

Scherer, R., & Tiemann, R. (2012). Factors of problem-solving competency in a virtual chemistry environment: The role of metacognitive knowledge about strategies. *Computers & Education, 59*(4), 1199–1214.

Schunn, C. D., & Klahr, D. (1995). A 4-space model of scientific discovery. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 106–111). Hillsdale, NJ: Erlbaum.

Schunn, C. D., & Klahr, D. (2000). Discovery processes in a more complex task. In D. Klahr (Ed.), *Exploring science*: *The cognition and development of discovery processes* (pp. 161–199). Cambridge, MA: MIT Press.

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning & Individual Differences, 24,* 42–52.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7*(2), 268–288.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18*(9), 8–14.

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

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

Sternberg, R. J., & Berg, C. A. (Eds.). (1992). *Intellectual development*. Cambridge: Cambridge University Press.

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

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

Süß, H. M. (1999). Intelligenz und komplexes problemlösen [Intelligence and complex problem solving]. *Psychologische Rundschau, 50*(4), 220–228.

Süß, H.-M. (2001). Prädikative Validität der Intelligenz im schulischen und außerschulischen Bereich. In E. Stern & J. Guthke (Eds.), *Perspektiven der Intelligenzforschung* [Perspectives on intelligence research] (pp. 109–135). Lengerich: Pabst Science.

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

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

Wagener, D. (2001). *Psychologische diagnostik mit komplexen szenarios* [Psychological diagnostic using complex scenarios]. Lengerich: Pabst.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology, 23*(1), 63–71.

Weaver, W. (1948). Science and complexity. *American Scientist, 36*, 536–544.

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

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

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

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*(4), 393–409.

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

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

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning, 19*, 127–146.

3

Complex problem solving and intelligence: A meta-analysis

This article is available as:

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92-101.

Intelligence 53 (2015) 92–101

Contents lists available at ScienceDirect

Intelligence

Complex problem solving and intelligence: A meta-analysis☆

Matthias Stadler [a,*], Nicolas Becker [b], Markus Gödker [a], Detlev Leutner [c], Samuel Greiff [a]

[a] University of Luxembourg, Luxembourg
[b] Saarland University, Germany
[c] Duisburg-Essen University, Germany

ARTICLE INFO

Article history:
Received 29 May 2015
Received in revised form 17 August 2015
Accepted 28 September 2015
Available online xxxx

Keywords:
Complex problem-solving
Dynamic decision making
Intelligence
Meta-analysis
Multiple complex systems

ABSTRACT

The purpose of this meta-analysis is to examine the nature and magnitude of the relation between complex problem-solving skills (CPS) and intelligence, a topic that has been widely discussed and that has instigated a vast array of partially contradicting findings in the past. Theoretically, researchers have hypothesized the two constructs to be everything from completely separate to identical. Over the course of almost four decades, empirical studies yielded results in support of both arguments. Our meta-analysis of 47 studies containing 60 independent samples and a total sample size of 13,740 participants revealed a substantial correlation of CPS and intelligence with an average effect size of $M(g) = .433$. In addition, we investigated whether the operationalization of CPS and intelligence moderated this correlation. Whereas there were no significant correlation differences considering the operationalization of intelligence, the approach used to measure CPS moderated the correlation of CPS and intelligence. Especially the most recent approach towards the assessment of CPS yielded the strongest associations between the two constructs. Implications for existing theories and future research are discussed.

1. Introduction

As the complexity and interconnectedness of the systems that we interact with in our daily lives increases, so does the importance of research on how we learn to control such complex environments. Just dealing with everyday objects (e.g., phones, computers, automated driving systems) requires being aware of their respective connections to other objects or people (e.g., other computers and people via the internet) as well as the inner workings of the objects themselves. In response to this growing challenge, Dörner and Kreuzig (1983) introduced the research area of complex problem solving (CPS) that focused on the assessment of individuals' ability to deal with complex and dynamically changing environments. This then promising new approach towards human ability was primarily brought forward by German researchers who were interested in experimentally investigating the interindividual differences among people's ability to solve complex simulations of real-world problems. In that, the assessment of CPS was considered a more ecologically valid alternative to established measures of human ability such as general intelligence. Especially initial theoretical propositions (e.g., Dörner & Kreuzig, 1983) and empirical findings (e.g., Putz-Osterloh, 1981) in favor of a clear distinction from general intelligence soon resulted in a plethora of research on the relation between the two constructs (e.g. Beckmann & Guthke, 1995; Wittmann & Süß, 1999). the results of these studies repeatedly contradicted each other with researchers hypothesizing and finding diverse results ranging from non-significant (e.g. Joslyn & Hunt, 1998; Putz-Osterloh, 1981) to very strong correlations between measures of CPS and general intelligence (e.g. Funke & Frensch, 2007, Wirth & Klieme, 2003, Wüstenberg, Greiff, & Funke, 2012). These results were in effect interpreted as either support for the discriminant validity of CPS as a construct or evidence that measures of CPS were actually measuring nothing else than general intelligence.

The purpose of the present work is, therefore, to answer the question on the empirical relation between CPS and intelligence by meta-analytically summarizing the various research findings on the correlation of CPS and intelligence. In addition, we will try to find moderating factors that might help explain the contradicting findings. Showing that there is a substantial but far from perfect correlation between various different measures of CPS and intelligence, we provide important information on the construct validity and nomological classification of CPS. Furthermore, our study investigates the moderating effects of different operationalizations suggesting differences in the assessment of CPS to be a potential explanation for the variation in results.

☆ This research was funded by grants of the Fonds National de la Recherche Luxembourg (ATTRACT, "ASKI21"; AFR "CoPUS").

* Corresponding author at: University of Luxembourg, Maison des Sciences Humaines, 11 Porte des Sciences à Esch-Belval L-4366, Luxembourg.
E-mail address: matthias.stadler@uni.lu (M. Stadler).

http://dx.doi.org/10.1016/j.intell.2015.09.005
0160-2896/

2. Complex problem solving and intelligence

Following a definition by Buchner (according to Frensch & Funke, 1995, p. 14), CPS is throughout this paper understood as:

> "(…) *the successful interaction with task environments that are dynamic (i.e., change as a function of the user's interventions 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.*"[1].

Considering this definition, it becomes obvious why CPS has often been compared to intelligence on a conceptual basis (Funke & Frensch, 2007) to establish discriminant validity or to characterize individual abilities that would help explain performance in CPS tasks. On the one hand, some characteristic features of CPS such as the integration of information are part of almost every definition of intelligence (Sternberg & Berg, 1986). On the other hand, the dynamic and intransparent characteristics of complex problems are not established aspects of the current conceptions of intelligence such as the Cattell–Horn–Caroll (CHC) theory of human intelligence (McGrew, 2009), and this aspect of CPS may, thus, be an important addition to the understanding of human ability (Dörner & Kreuzig, 1983; Greiff et al., 2013).

This theoretical ambiguity is reflected in empirical findings on the relation between CPS and intelligence. Multiple early studies indicated that, while performance in CPS tasks varied tremendously among individuals, psychological assessments of general intelligence were unable to explain this variability (Brehmer, 1992; Rigas & Brehmer, 1999). Kluwe, Misiak, and Haider (1991) summarized 11 of these early studies on the relation between CPS and intelligence and concluded that most of them failed to show a close relation between intelligence scores and CPS performance measures. This led several researchers to suggest CPS to be a cognitive construct mostly independent from intelligence (Putz-Osterloh, 1985). Rigas and Brehmer (1999) summarized this view in the *different-demands hypothesis*. To explain the weak correlations that researchers observed between measures of general intelligence and CPS performance, this hypothesis suggests that CPS tasks demand the performance of more complex mental processes than intelligence tests do, such as the active interaction with the problem to acquire knowledge on the problem environment, which, in turn, results in low empirical correlations between the constructs.

Whereas there is some support for the different-demands hypothesis (e.g., Joslyn & Hunt, 1998), more recent studies challenge it. In a comprehensive study, Gonzalez, Thomas, and Vanyukov (2005) found correlations ranging from $r = .33$ to $r = .63$ between various measures of CPS and measures of general intelligence. Similarly, Süß, Kersting, and Oberauer (1991) reported correlations of $r = .40$ between *Tailorshop* performance measures (*Tailorshop* being one of the most frequently used measures of CPS) and measures of general intelligence.

Based on these moderate to strong correlations and contradicting initial assumptions of independence of the two constructs as put forward in the different-demands hypothesis, several researchers even argued that measures of CPS would be almost redundant to measures of general intelligence (Mayer et al., 2013; Wittmann & Süß, 1999). Wirth and Klieme (2003) reported a correlation of .84 between a latent factor of different measures of CPS and general intelligence. Similarly, latent factor scores on *MultiFlux*, a more recently developed measure of CPS (Kröner, 2001), showed a latent correlation of .75 with different facets of the Berlin Model of Intelligence Structure (BIS) test (Jäger, Süß, & Beauducel, 1997) an established intelligence test (Kröner, Plass, & Leutner, 2005).

The latest studies on the relation between CPS and intelligence also reported moderate to strong latent correlations (between $r = .50$ and $r = .80$) of the two constructs (e.g. Greiff et al., 2013; Sonnleitner et al., 2012; Wüstenberg, Stadler, Hautamäki, & Greiff, 2014; Wüstenberg et al., 2012). However, these studies additionally demonstrated incremental value over and above measures of intelligence in predicting school grades (Wüstenberg et al., 2012) and job success (Danner, 2011) despite these strong correlations and in support of the *different-demands hypothesis*.

An explanation for these inconsistent findings regarding the relation between CPS and intelligence may lie in the conceptualization of intelligence. Almost all current theories of psychometric intelligence include one or two very broad, latent factors of general intelligence that capture a large proportion of all cognitive abilities such as abstract reasoning, memory, or factual knowledge (McGrew, 2009). Based on this concept, early studies on the relation between CPS and intelligence mostly included rather broad measures of general intelligence (e.g., Putz-Osterloh, 1985) using different tasks assessing various cognitive abilities including factual knowledge (or general crystalized intelligence; McGrew, 2009). More recent studies, on the other hand, focused on specific sub-facets of intelligence, and especially reasoning was theoretically and empirically determined to be conceptually closest to CPS (Wittmann & Süß, 1999). Reconsidering the *different-demands hypothesis*, broad measures of intelligence may be covering several aspects that are not relevant for the successful solution to a complex problem, such as factual knowledge, thus limiting the empirical relation between CPS and intelligence. However, assessments focusing on reasoning (e.g., Raven's Progressive Matrices; Raven, Raven, & De Lemos, 1958) as "the use of deliberate and controlled mental operations to solve novel problems that cannot be performed automatically" (McGrew, 2009) are conceptually closer to CPS than very broad measures of general intelligence and may thus yield much stronger correlations of CPS and intelligence (e.g.Greiff et al., 2013, Wittmann & Hattrup, 2004, Wittmann & Süß, 1999). Accordingly, the conceptualization of intelligence used in a study may influence the relation between CPS and intelligence found with higher correlations of CPS and reasoning than of CPS and broad measures of general intelligence.

3. Assessment of complex problem solving

In the same way, the assessment approach used to measure CPS varied greatly among studies and may be responsible for the variation in findings on the relation between CPS and intelligence. The assessment of abilities such as CPS entails by definition (Frensch & Funke, 1995) the possibility of an active interaction between the person to be assessed and the assessment instrument. As no such interaction is possible within paper-pencil tests, this necessarily requires a computer-based assessment. With the advancement of computer technology, various CPS tasks have evolved following different approaches. Next to different conceptualizations of intelligence, this diversity in assessment approaches for CPS may be another cause for the inconsistent results regarding the relation between CPS and intelligence.

The first computer-based CPS tasks were developed in the early 1980s with the aim of administering task environments with a high resemblance to the real world and the goal of producing a reliable and ecological valid measure of CPS that sufficiently simulated reality. The microworld *Lohausen* (Dörner, Kreuzig, Reither, & Stäudel, 1983), for example, required a participant to govern a small city, which was intricately simulated with more than 1000 different and interconnected variables. Whereas these *classical measures of CPS* enjoyed a high level of face validity, their psychometric properties were rather problematic (Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015). Measurement issues, such as unsatisfactory reliability and validity, quickly raised doubt on the measurability and validity of the construct of CPS itself

[1] Several other constructs describing the ability of dealing with complex environments have been suggested to extend the existing host of human abilities. Most prominent among those are Dynamic Decision Making (DDM; Brehmer, 1992) and Systems Thinking (Booth-Sweeney & Sterman, 2000). Both of these constructs overlap greatly with CPS in their respective definitions (Frensch & Funke, 1995), and the variation in terminology is mostly due to different research traditions. Throughout this paper, we will focus primarily on CPS but also refer to relevant results published under different labels.

(Kröner et al., 2005; Wüstenberg et al., 2012). Rigas, Carling, and Brehmer (2002) summarized these problems in suggesting the *low-reliability hypothesis* to explain why prior research failed to establish an association between performance in CPS tasks and intelligence. In fact, there is convincing evidence that the poor reliability of some *classical measures* employed in past studies made it difficult to find any relations to other constructs at all (for an overview see Greiff, 2012; Rigas et al., 2002).

In reaction to these problems, Funke (2001) introduced Linear Structural Equation systems (LSE) and Finite State Automata (FSA) as formal frameworks that allow for the description of underlying task structures. Both of these frameworks enabled the creation of *single complex systems*, which are independent of any semantic embedment (Greiff, Fischer, Stadler, & Wüstenberg, 2014) as they only specify an underlying system that can be clad in multiple semantic contexts.

In particular, the LSE formalism has been widely adopted by CPS research and has led to the development of a considerable number of *single complex systems* (e.g., *Multiflux*, Kröner, 2001; *FSYS*, Wagener, 2001). In a further advancement, after Leutner, Klieme, Meyer, and Wirth (2004) had used a combination of two single complex systems for measuring CPS, Greiff, Wüstenberg, and Funke (2012) used the LSE framework for the development of the *multiple complex systems* (MCS; Greiff et al., 2014) approach, which was featured in the Program for International Student Assessment (PISA) 2012, the arguably most important large-scale assessment worldwide. This approach solves several measurement issues by using multiple small rather than one single, large microworld as in *classical measures* of CPS or *single complex systems* relying on LSE or FSA (Greiff et al., 2014). This approach was realized in assessment tools such as *MicroDYN* (Greiff et al., 2012) or *Genetics Lab* (Sonnleitner et al., 2012) and was later extended to FSA with the development of MicroFIN (Neubert, Kretzschmar, Wüstenberg, & Greiff, 2014). These *MCS* measures of CPS were developed with a clear focus on psychometric quality and showed significantly higher reliability than classical measures of CPS. In concordance with the *different-demands hypothesis*, they were also found to correlate substantially with measures of intelligence (e.g. Sonnleitner et al., 2012; Wüstenberg et al., 2012).

An alternative explanation for the fuzzy results of studies on the relation of intelligence and CPS could lie in the semantic embedment of CPS tasks. The *Elshout–Raaheim hypothesis* (Elshout, 1987; Raaheim, 1988; see also Leutner, 2002) proposes an inverted U-shaped relation between the score of the correlation coefficient as the dependent variable and the amount of available domain-specific knowledge as the independent variable. As *classical measures* of CPS emulated real-world problems, domain specific knowledge could be used to solve the problems, thus limiting the relevance of individual intelligence. More recent measures of CPS based on LSE or FSA (both *single complex systems* and *MCS*) are less dependent of a semantic context, and, thus, less domain specific knowledge can be used. This should result in a stronger relation between performance in modern CPS tasks and intelligence.

In summary, following both the *low-reliability hypothesis* (Rigas et al., 2002) and the *Elshout–Raaheim hypothesis* (Leutner, 2002), the approach used to assess CPS in different studies could moderate the relation between CPS and intelligence.

4. The present research

Based on the wide range of research with partially contradicting findings on the relation between CPS and intelligence presented above, it seems necessary to meta-analytically summarize these findings for the first time ever.

In addition, two possible explanations for the contradicting results regarding the correlation of CPS and intelligence seem to be plausible. One the one hand, it may be necessary to differentiate between studies that employed very broad measures of general intelligence capturing multiple sub-facets and those that focused on more specific sub-facets such as reasoning (Wittmann & Süß, 1999). Whereas earlier studies that found small correlations predominantly considered more general measures of intelligence (for a summary see Kluwe et al., 1991; Beckmann, 1994), more recent studies, focusing on reasoning, consistently report higher correlations of CPS and reasoning (e.g.Danner, 2011, Greiff et al., 2013, Greiff et al., 2014, Sonnleitner et al., 2012). Thus, we will investigate whether the difference in operationalization of intelligence moderates the relation of CPS and intelligence.

On the other hand, advancements in CPS measurement may have increased the reliability and reduced the semantic embedment of CPS assessment instruments, thus theoretically allowing for higher correlations with other measures (Leutner, 2002; Rigas et al., 2002). Therefore the second moderator investigated in this study will be the operationalization of CPS.

In summary, the present research will meta-analytically summarize the empirical findings available to answer the question on the relation between CPS and intelligence. In a second step, we will investigate whether the conceptualization of intelligence (measures of general intelligence or measures of reasoning) or the conceptualization of CPS (classical measures of CPS, single complex systems, or MCS tests) can be used to explain the variation among those findings.

5. Method

5.1. Literature search

5.1.1. Compilation of database

We used three strategies to identify studies for the present meta-analysis: (1) We conducted a broad literature search using the databases PsycINFO, PsycARTICLES, and PSYNDEX. Search terms for intelligence were “Reasoning”, “Intelligence”, “Working Memory”, “Short-Term Memory”, and “Reaction-Time”. Search terms for CPS were “Complex Problem Solving”, “Dynamic Problem Solving”, “Interactive Problem Solving”, “Microworlds”, “Systems Thinking”, and “Dynamic Decision Making”. The search terms were combined in all 30 [5(Intelligence) × 6(Complex Problem Solving)] possible ways. Entering these combinations in the 3 databases resulted in 90 queries. (2) We conducted an additional unsystematic search of literature based on publications of well-known authors within the fields of CPS and intelligence as well as on publications referenced in those publications. The systematic and unsystematic search resulted in 123 different studies, which seemed relevant according to the title and abstract. (3) As research on CPS was primarily brought forward by German researchers, we contacted the mailing list of the “Deutsche Gesellschaft für Psychologie”, the German Association of Psychology, to gather “gray literature” and reduce publication bias. We asked the members to send us information about unpublished studies yielding correlations between intelligence and complex problem solving. This appeal resulted in 7 additional datasets to be considered.

5.1.2. Inclusion criteria

In this meta-analysis, we considered all studies that fulfilled the following inclusion criteria: (1) Intelligence was measured by a standardized intelligence test; (2) CPS was measured by a standardized complex scenario; (3) the study reported zero-order correlations of intelligence and CPS or a coefficient that allowed the calculation of a zero-order correlation; (4) the sample size of the study was reported.

5.1.3. Exclusion of studies

The total amount of 130 studies was checked whether they fulfilled the inclusion criteria or not. 7 studies (5.55%) did not use a standardized CPS measurement, and 7 studies (5.55%) were excluded because they did not assess intelligence by a standardized intelligence test. 9 studies (7.14%) did not report any correlations of the CPS and intelligence measures. 61 studies (48.41%) were excluded because they did not report an empirical study or because they reported results from studies that had

Table 1
Description and effect size estimates for all independent samples included in the meta-analysis.

ID	Author(s)	Year	N	CPS measure	Intelligence measure	r
1	Abele et al.	2012	167	MCS	Reasoning	.40
2	Beckman & Guthke	1995	92	Classical	General intelligence	.15
3	Bühner et al.	2008	144	SCS	Reasoning	.16
4	Burkolter et al.	2009	41	Classical	General intelligence	.75
5	Burkolter et al.	2010	39	Classical	General intelligence	.22
6	Burmeister	2009	44	Classical	General intelligence	.47
7	Danner	2011	173	SCS	Reasoning	.86
8	Dörner et al. Sample 1	1983	48	Classical	Reasoning	−.03
9	Dörner et al. Sample 2	1983	48	Classical	Reasoning	.12
10	Gediga et al.	1984	29	Classical	General intelligence	.09
11	Gonzales et al. Sample 1	2005	15	Classical	Reasoning	.71
12	Gonzales et al. Sample 2	2005	28	Classical	Reasoning	.63
13	Gonzales et al. Sample 3	2005	74	Classical	Reasoning	.33
14	Greiff & Fischer	2013	140	MCS	Reasoning	.50
15	Greiff et al. Sample 1	2015	339	Classical	Reasoning	.24
16	Greiff et al. Sample 2	2015	339	MCS	Reasoning	.52
17	Güss & Dörner	2011	511	Classical	General intelligence	.19
18	Hasselmann	1993	21	Classical	General intelligence	.26
19	Hesse Sample 1	1982	30	Classical	Reasoning	−.17
20	Hesse Sample 2	1982	30	Classical	Reasoning	.06
21	Hesse Sample 3	1982	30	Classical	Reasoning	.38
22	Hesse Sample 4	1982	30	Classical	Reasoning	.46
23	Hörmann & Thomas Sample 1	1989	19	Classical	General intelligence	.46
24	Hörmann & Thomas Sample 2	1989	21	Classical	General intelligence	−.03
25	Hussy Sample 1	1985	15	Classical	Reasoning	−.30
26	Hussy Sample 2	1985	15	Classical	Reasoning	.25
27	Hussy Sample 3	1985	15	Classical	Reasoning	.35
28	Hussy Sample 4	1985	15	Classical	Reasoning	.50
29	Hussy	1989	154	Classical	General intelligence	.38
30	Kersting	2001	99	Classical	General intelligence	.26
31	Klieme et al.	2001	650	Classical	Reasoning	.58
32	Kluge et al.	2011	38	Classical	General intelligence	.13
33	Kretzschmar	2010	118	SCS	General intelligence	.30
34	Kretzschmar et al.	Unpublished	197	MCS	General intelligence	.34
35	Kröner	2001	28	SCS	Reasoning	.51
36	Kröner et al.	2005	101	SCS	Reasoning	.67
37	Leutner et al.	2004	535	MCS	General intelligence	.63
38	Leutner et al.	2005	654	Classical	Reasoning	.84
39	Leutner Sample 1	2002	200	Classical	Reasoning	.43
40	Leutner Sample 2	2002	28	Classical	Reasoning	.05
41	Neubert et al.	2014	576	MCS	Reasoning	.62
42	Putz-Osterloh	1985	50	Classical	General intelligence	.36
43	Rigas et al.	2002	62	Classical	Reasoning	.33
44	Scherer & Tiemann a	2014	805	SCS	Reasoning	.55
45	Scherer & Tiemann b	2014	1487	SCS	Reasoning	.58
46	Sonnleitner et al.	2012	61	MCS	Reasoning	.30
47	Stadler et al.	In press	78	SCS	General intelligence	.20
48	Stadler et al. Sample 1	Unpublished	161	MCS	Reasoning	.83
49	Stadler et al. Sample 2	Unpublished	254	MCS	Reasoning	.74
50	Süß et al.	1991	127	Classical	Reasoning	.47
51	Süß et al.	1993	214	Classical	Reasoning	.40
52	Wagener & Wittmann	2002	35	SCS	Reasoning	.63
53	Wagener Sample 1	2001	63	SCS	Reasoning	.31
54	Wagener Sample 2	2001	71	SCS	General intelligence	.20
55	Wagener Sample 3	2001	136	SCS	General intelligence	.47
56	Wagener Sample 4	2001	51	SCS	General intelligence	.24
57	Wirth & Funke	2005	688	SCS	Reasoning	.46
58	Wittmann et al.	1996	92	Classical	General intelligence	.57
59	Wüstenberg et al.	2012	222	MCS	Reasoning	.59
60	Wüstenberg et al.	2014	3191	MCS	Reasoning	.66

Note. CPS = Complex problem solving; r = correlation coefficient; MCS = Multiple complex systems; SCS = Single complex systems.

already been reported in another study. In the end, a total amount of 47 studies containing 60 independent samples were included in the meta-analysis (Table 1).

5.1.4. Coding of measures

The measures used in the final set of studies were coded according to the hypothesized moderators' levels by two independent raters. Measures of CPS were coded as either classical measures of CPS, single complex systems (SCS), or MCS tests. Measures of intelligence were defined as either measures of general intelligence or measures of reasoning. This classification was a clear and unambiguous task resulting in a perfect agreement between the two raters. Table 2 displays the coding for all measures of CPS and intelligence used in the studies included into the meta-analysis.

5.2. Meta-analytic procedure

5.2.1. Main meta-analysis

In our analysis we followed the guidelines described by Field and Gillett (2010) and used the SPSS 20.0.0 (IBM, 2011) and R 3.2.1 (R Core Team, 2015) syntaxes provided there. We chose to employ a random-effects model because it can be assumed that the true effects

Table 2
Coding for the measures of CPS and intelligence.

CPS measures		
Classical	SCS	MCS
AGRIMAN	Chemie Labor [Chemistry Lab]	Genetics Lab
Cabin Air Management System	Heidelberg Finite State Automaton [Space Shuttle]	Schmetterlings−/Parabelproblem [Butterfly/Parabola Problem]
Cherry-Tree	FSYS	MicroDYN
Coldstore	K4	MicroFIN
DISKO	MultiFlux	
Dori (Sahel)	M3	
Firechief		
Hamurabi		
Hunger in the Sahel		
Learn		
Lohhausen		
Moro		
Powerplant		
Tailorshop		
Textilfabrik		
Water Purification Plant		
WinFIRE		

Intelligence measures	
General intelligence	Reasoning
Intelligenz Struktur Analyze (ISA)	Culture Fair Test (CFT) 20-R
Leistungsprüfsystem (LPS)	Cognitive ability test (CogAT)
Berliner Intelligenz Strukturtest (BIS)	Standard Progressive Matrices (SPM)
Intelligenz Struktur Test (IST)	Advanced Progressive Matrices (APM)
Wonderlic Personnel Test	IST — Subtests (Figures, Dices, Matrices, Analogies)
Leistungsprüfsystem (LPS)	BIS-K
	Kognitiver Fähigkeitstest (KFT) — figural scale

Note. CPS = Complex problem solving; MCS = Multiple complex systems; SCS = Single complex systems; only measures for which names were provided in the manuscripts are listed.

vary between the studies (e.g., due to different conceptualizations of CPS and intelligence). Furthermore, we chose to employ the meta-analytic strategy of Hedges and Vevea (1998) rather than the strategy of Hunter and Schmidt (2004) because the 95% confidence intervals of the latter one tend to be too small whereas both strategies provide comparably accurate estimates of the population effect size (see Field, 2005). As some of the studies employed yield rather small sample sizes, the correlations were converted to Hedges' g prior to conducting the meta-analysis (Hedges, 1981). We computed the mean weighted Hedges' g [M(g)] as an estimate of the population effect size, the associated 95% confidence bounds (95% CI_u; 95% CI_l) as an indicator of the significance of the population effect, the estimated variance in the population (τ^2) as an indicator of the variability of the effects in the population, and the Q statistic as an indicator of the homogeneity of effect sizes. Additionally, we computed I^2 (Borenstein, Hedges, Higgins, & Rothstein, 2009; Higgins, Thompson, Deeks, & Altman, 2003) as an additional heterogeneity estimate, which describes what proportion of the observed variance reflects real differences in effect size (signal-to-noise ratio). I^2 values can range between 0% and 100%. Values on the order of 25%, 50% and 75% are considered as low, moderate and high, respectively.

5.2.2. Outlier and influence analyses

Outliers and influential cases were identified using the package metafor (Viechtbauer, 2010) in R 3.2.1 (R Core Team, 2015) following the guidelines of Viechtbauer and Cheung (2010). Outliers were identified by computing standardized deleted residuals (SDRs) for each study, which represent the deviation of the correlation of a single study from the mean correlation of all other studies expressed in standard deviations. Studies with SDRs above 1.96 or below −1.96 were regarded as substantial outliers. To analyze the influence of outliers on the mean correlation of the meta-analysis, we computed Cook's distance (CD) and COVRATIO values for each study. CD can be interpreted as the Mahalanobis distance between the predicted average correlation for the study once with and once without the study included in the model fitting. Following Cook and Weisberg (1982) we regarded studies with CD values greater than .45 as having a substantial influence on the main effect. The COVRATIO of a study describes the change of the variance-covariance matrix of the parameter estimates when the study is excluded. Viechtbauer and Cheung (2010) view COVRATIOs smaller than 1 as an indicator, that the exclusion of the concerned study improves the precision of the model parameters. Furthermore, we computed the meta-analysis with and without the outliers to provide a direct comparison of the results with and without outliers.

5.2.3. Moderator analyses

For testing moderator effects we applied random-effects regression analysis as recommended by Field and Gillett (2010). In this analysis a general linear model is assumed in which the effect sizes are predicted as a function of the moderator variable. The significance of the moderator effect can be assessed using a χ^2-statistic (for further information see Field, 2003; Overton, 1998). Furthermore we computed Q-tests for subgroup heterogeneity as recommended by Borenstein et al. (2009) and regarded significant Q-values between groups (Q_{bet}) as an indicator of a moderating effect.

5.2.4. Identification of publication bias

As our analyses rely predominantly on published studies the possibility of publication bias had to be considered. Publication bias refers to the fact that significant results are more likely to be published than insignificant ones, what might lead to an overestimation of the effects found in meta-analyses. In order to identify possible publication bias in the present study, we analyzed the association between Hedges' g and standard errors using Kendall's τ. As recommended by Begg and Mazumdar (1994), a significant Kendall's τ value can be interpreted as indicator of publication bias. The results of these analyses showed a slight publication bias (see results section). We therefore corrected the results for publication bias using the strategy of Vevea and Woods (2005) who suggest modeling the likelihood of a study being published according to their weights. The mean Hedges' g corrected for moderate publication bias [$M(g)_{corr}$] and population variance (τ^2_{corr}) was compared with the mean correlation of the initial meta-analysis in order to assess the effect of publication bias on the results of this study.

Table 3
Stem and leaf display of effect sizes (r) from 59 samples.

Stem	Leaf
.8	3, 4, 6
.7	1, 4, 5
.6	2, 3, 3, 3, 6, 7
.5	0, 0, 1, 2, 5, 7, 8, 8, 9
.4	0, 0, 3, 6, 6, 6, 7, 7, 7
.3	0, 0, 1, 3, 3, 4, 5, 6, 8, 8
.2	0, 0, 2, 4, 4, 5, 6, 6
.1	2, 3, 5, 6, 9
.0	5, 6, 9
−.0	3, 3
−.1	7
−.2	
−.3	0

Note. If a sample had more than one effect size, the mean effect size was calculated and is reported in the table.

6. Results

6.1. Meta-analysis of all studies

Table 3 displays the effect sizes found for all studies in a stem-and-leaf plot. As can be seen, there was a wide range of correlation coefficients ranging from r = −.30 to r = .86.

The results of the meta-analysis of all studies are presented in Table 4. The mean weighted Hedges' g was M(g) = .433 and the population variance was $\tau^2 = .071$. As the 95% confidence interval ranged from .370 to .492 the mean Hedges' g of complex problem solving and intelligence could be regarded as significantly greater than zero. The homogeneity of the distribution of Hedges' g values could be assumed since the Q-statistic was not significant (p = .228). An I^2-value of 93.7% indicated that the observed variance almost exclusively reflects real differences in effect size.

To investigate how robust this finding was, considering different levels of reliability for our measures of CPS and intelligence, we conducted a sensitivity analysis correcting for unreliability under a range of reliability assumptions. The results of this analysis are displayed in Table 5. As can be seen, the mean effect size did not exceed a Hedge's g of M(g) = .607 even when poor reliabilities ($r_{xx} = .60$) were assumed for both measures of CPS and intelligence. This confirms the interpretation of CPS and intelligence as highly related but separable constructs.

Because it is possible to find moderating effects although the distribution of effect sizes is homogeneous (Hall & Rosenthal, 1991), we decided to additionally conduct meta-analyses on moderator levels.

6.2. Outlier and influence analyses

Fig. 1 presents the results of the outlier and influence analyses. It can be recognized, that except for three studies (5, 28, 57) the SDRs did not exceed the cut-off value for substantial outliers. Thus, the Hedges' g values of the studies employed in the meta-analysis can be regarded as rather homogeneous. The CD values were below the cut-off for all studies. The COVRATIO values were substantially below the cut-off only for the studies, which were identified as outliers. The results of the meta-analysis without outliers can be found in Table 4. It can be recognized that the mean weighted Hedges' g values [M(g) = .433 vs. M(g) = .399] as well as population variances ($\tau^2 = .071$ vs. $\tau^2 = .046$) did not differ substantially from another. Furthermore there was substantial overlap between the 95% confidence intervals of the two analyses [.370 ≤ M(g) ≤ .492 vs. .343 ≤ M(g) ≤ .453]. Therefore we concluded, that the results of the main meta-analysis are rather robust against outliers.

6.3. Moderator analyses

The results of the moderator analyses are presented in Table 4. For studies operationalizing intelligence by reasoning tests, the mean weighted Hedges' g was M(g) = .472 with a population variance of $\tau^2 = .064$ and a 95% confidence interval ranging from .400 to .538. For studies operationalizing intelligence by measures of general intelligence, the mean weighted Hedges' g was M(g) = .360 with a population variance of $\tau^2 = .052$ and a 95% confidence interval ranging from .257 to .454. The results of the random effects regression analysis [$\chi^2(1) = 3.206$; p = .073] indicated that the moderating effect of the operationalization of intelligence cannot be regarded as significant. This was supported by the result of a Q-test for subgroup heterogeneity which proved as insignificant [$Q_{bet}(1) = 3.406$; p = .182].

For studies operationalizing CPS by classical measures of CPS, the mean weighted Hedges' g was M(g) = .339 with a population variance of $\tau^2 = .142$ and a 95% confidence interval ranging from .213 to .454. Studies in which CPS was operationalized by single systems based on LSE (SCS), showed a mean weighted Hedges' g of M(g) = .471 with a population variance of $\tau^2 = .051$ and a 95% confidence interval ranging from .363 to .566. For studies operationalizing CPS by MCS tests, the mean weighted Hedges' g was M(g) = .585 with a population variance of $\tau^2 = .029$ and a 95% confidence interval ranging from .510 to .652. The results of the random effects regression analysis [$\chi^2(2) = 9.620$; p = .008] indicated that the moderating effect of the operationalization of CPS can be regarded as significant. This result was supported by the result of a Q-test for subgroup heterogeneity, which proved as significant [$Q_{bet}(2) = 12.984$; p = .002].

Moreover, the corrected variance (τ^2) within the studies using classical measures of CPS was larger than the corrected variance within all studies. To further investigate this unexpected result, we conducted an additional interaction analysis separating the studies using classical measures of CPS and measures of general intelligence from those using classical measures of CPS and measures of reasoning. The result of this interaction analysis can be found in Table 4. The average effect sizes did not differ significantly between both subgroups [$Q_{bet}(1) = .075$; p = .963] but the corrected variance for studies using classical measures of CPS and measures of general intelligence fell below the corrected variance within all studies ($\tau^2 = .031$) whereas the corrected variance for studies using classical measures of CPS and measures of reasoning remained higher ($\tau^2 = .166$).

Finally, we repeated our moderation analysis assuming plausible values as average reliability coefficients for each type of CPS measure to investigate whether the moderating effect of different operationalizations of CPS on the correlation of CPS and intelligence is due to different levels of reliability. We assumed poor reliability ($r_{xx} = .60$) for classical measures of CPS, very good reliability for SCS

Table 4
Meta-analytic results and moderator analyses.

Analysis	*k*	*M(g)*	τ^2	*95% CI$_l$*	*95% CI$_u$*	*Q*	*df*	*p*	I^2
All Studies	60	.433	.071	.370	.492	66.763	59	.228	93.700%
Without outliers	57	.399	.046	.343	.453	58.399	56	.387	90.306%
Measure of intelligence									
Reasoning	39	.472	.064	.400	.538	56.247	38	.029	94.299%
General intelligence	21	.360	.052	.257	.454	15.587	20	.742	84.214%
Measure of CPS									
MCS	11	.585	.029	.510	.652	18.870	10	.042	91.782%
SCS	14	.471	.051	.363	.566	20.439	13	.085	92.294%
Classical	35	.339	.142	.213	.454	18.673	34	.985	93.320%
Interaction									
Classical × Reasoning	21	.351	.166	.174	.505	12.012	20	.961	94.804%
Classical × General Intelligence	14	.323	.031	.212	.426	14.513	13	.339	69.117%

Note. *k* = number of studies; M(g) = mean Hedges' g; τ^2 = estimated variance in population; CI$_l$ = lower bound of 95% confidence interval; CI$_u$ = upper bound of 95% confidence interval; Q = Q statistic; df = degrees of freedom of Q statistic; p = significance of Q; CPS = Complex problem solving; MCS = Multiple complex systems; SCS = Single complex systems.

Table 5
Sensitivity analysis for reliability.

	$r_{Int;Int} = .60$	$r_{Int;Int} = .70$	$r_{Int;Int} = .80$	$r_{Int;Int} = .90$
$r_{CPS;CPS} = .60$	M(g) = .607 Q = 42.151 p = .952	M(g) = .551 Q = 56.963 p = .551	M(g) = .595 Q = 46.459 p = .882	M(g) = .550 Q = 52.979 p = .696
$r_{CPS;CPS} = .70$		M(g) = .585 Q = 47.697 p = .854	M(g) = .576 Q = 71.900 p = .121	M(g) = .528 Q = 69.207 p = .171
$r_{CPS;CPS} = .80$			M(g) = .523 Q = 69.182 p = .171	M(g) = .521 Q = 48.707 p = .828
$r_{CPS;CPS} = .90$				M(g) = .491 Q = 65.247 p = .267

Note. $r_{Int;Int}$ = Reliability Intelligence; $r_{CPS;CPS}$ = Reliability Complex Problem Solving; M(g) = Mean weighted Hedges' g; Q = Q-value for heterogeneity; p = Significance.

measures ($r_{xx} = .85$) as well as MCS measures ($r_{xx} = .80$). These values are plausible based on the existing body of literature (Greiff et al., 2014) and are being discussed more thoroughly below. The results of this analysis are displayed in Table 6. The general pattern of effect sizes remains the same with an average of M(g) = .447 for classical measures, M(g) = .577 for SCS, and M(g) = .720 for MCS tests. This difference remained significant $Q_{bet}(2) = 13.208$ ($p < .001$).

6.4. Analysis of publication bias

Regarding publication bias, we found a significant Kendall's $\tau = -.274$ ($p = .003$) between the Hedges' g values and the respective standard errors. This indicated a slight publication bias in both ways. That is, the variance of the Hedges' g values was higher than what would be expected by chance. We therefore chose to additionally correct the mean weighted Hedges' g for moderate two-tailed selection following the guidelines of Vevea and Woods (2005). The mean corrected correlation showed a value of $M(r)_{corr} = .412$ with a corrected population variance of $\tau^2_{corr} = .087$. As these values did not differ substantially from values of the main meta-analysis [M(r) = .428; $\tau^2 = .084$] we regarded our results as relatively unaffected from publication bias.

7. Discussion

This meta-analysis investigated the relation between intelligence and complex problem solving (CPS). The findings show a substantial mean effect size of M(g) = .433 for the correlation of the two constructs that is highly significant with only little evidence for publication bias. This finding contradicts earlier reviews suggesting a non-significant relation between CPS and intelligence (Kluwe, 1991) and suggests that intelligent people also tend to be more successful in dealing with complex problem-solving tasks. On the other hand, the results do not support the

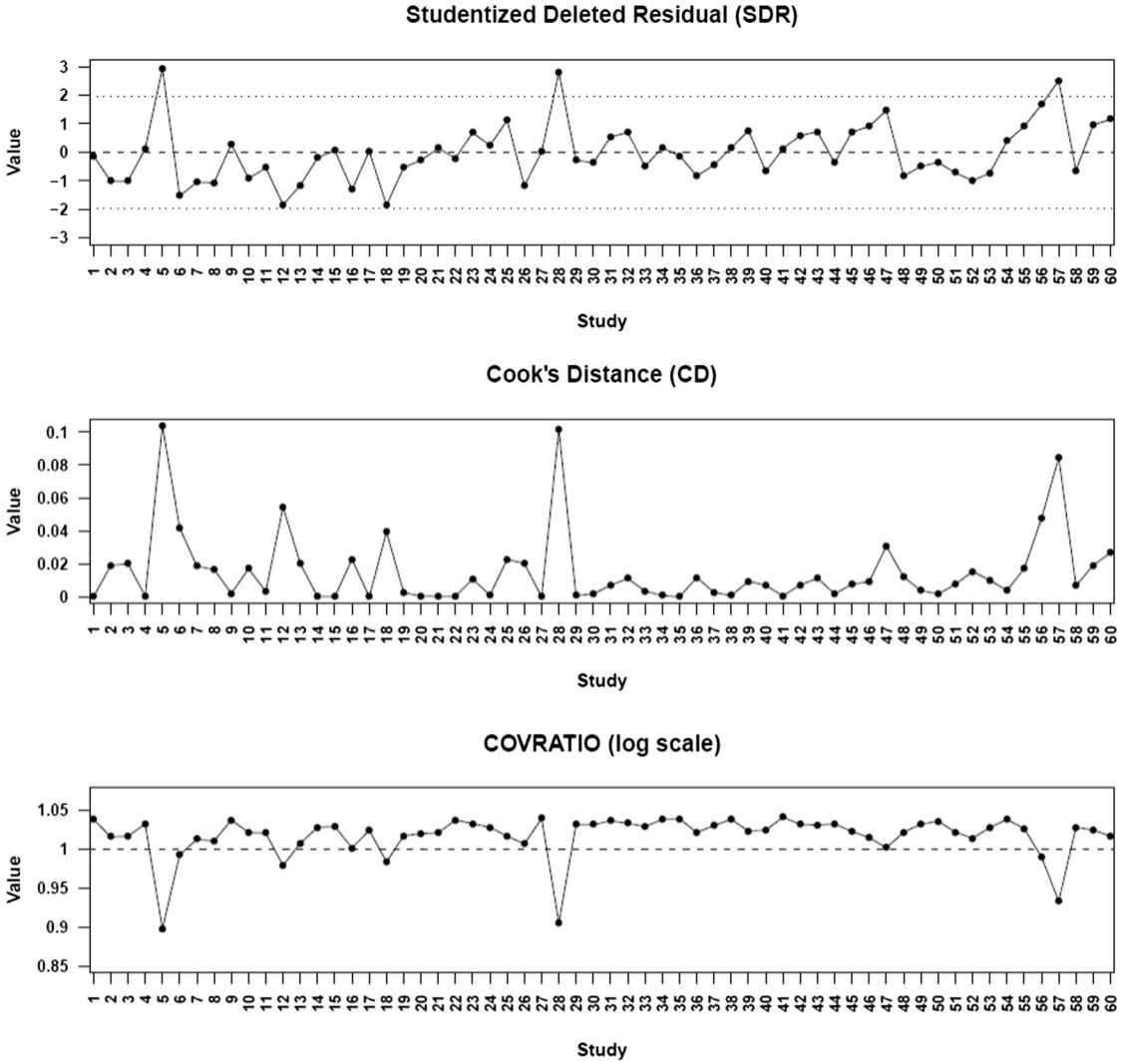

Fig. 1. Results of the outlier and influence analyses.

Table 6
Moderator analysis for measure of CPS corrected for reliability.

Analysis	*k*	*M(g)*	τ^2	*95% CI$_l$*	*95% CI$_u$*	*Q*	*df*	*p*	I^2
Measure of CPS									
MCS	11	.720	.106	.612	.802	22.598	10	.012	96.633%
SCS	14	.577	.175	.406	.708	20.395	13	.086	97.628%
Classical	35	.447	.112	.342	.540	36.963	34	.334	91.661%

Note. k = number of studies; M(g) = mean Hedges' g; τ^2 = estimated variance in population; CL_l = lower bound of 95% confidence interval; CL_u = upper bound of 95% confidence interval; Q = Q statistic; df = degrees of freedom of Q statistic; p = significance of Q; N_{FS} = Fail Safe N; CPS = Complex problem solving; MCS = Multiple complex systems; SCS = Single complex systems.

proposition of near to unity correlation of the two constructs either, as was discussed by several authors on the basis of single empirical studies (Kröner et al., 2005; Sonnleitner et al., 2012; Wittmann & Süß, 1999). In line with the *different-demands hypothesis* (Rigas & Brehmer, 1999), CPS performance could, thus, demand the enactment of more complex mental processes than do intelligence tests such as the active interaction with the problem to acquire knowledge on it.

The comprehensive answer to the question on the relation between CPS and intelligence however, appears to depend on the operationalization of CPS. Whereas the moderator analyses did not indicate significant differences between measures of general intelligence and measures of reasoning in respect to their relation to measures of CPS, there are substantial differences in mean effect sizes found for studies using different operationalizations of CPS. The smallest average effect size for the relation of CPS and intelligence was found for classical measures of CPS, M(g) = .339, followed by single systems based on LSE, M(g) = .471. CPS scores gained from MCS tests are related most strongly to intelligence, M(g) = .585.

Unexpectedly, the corrected variance (τ^2) within the studies using classical measures of CPS was larger than the corrected variance within all studies. This unexplained variance may be due to the effects of measures of intelligence within the classical distribution. There was no significant interaction effect between operationalization of intelligence and classical measures of CPS but only the corrected variance for studies using classical measures of CPS and measures of reasoning showed was found to be higher than the corrected variance of all studies. This suggests that there my be additional factors, such as the modality of reasoning tasks (e.g., figural vs. verbal) separating studies using classical measures of CPS and measures of reasoning from each other that were not included in this meta-analysis.

The significant moderator effect for operationalizations of CPS can be interpreted in three different ways. First, it appears to support the *low-reliability hypothesis* (Rigas et al., 2002) suggesting that unsatisfactory psychometric properties found for classical measures of CPS limit the correlation of CPS and intelligence. Reliability estimates for classical measures of CPS are rare (for an overview see Süß, 1996) and associated with several problems. Correlations between repeated measurements using the same classical measure of CPS are problematic as CPS is a process of active learning by interaction with the problem (Funke, 2001), resulting necessarily in knowledge about the task thus confounding any following assessment using the same measure (Wagener, 2001). On the other hand, the lack of a theoretical framework prohibits the creation of adequately parallel versions of a classical measure of CPS. The few estimates provided in the literature generally point towards a poor reliability ($r_{xx} < .70$) of classical measures of CPS (e.g. Rigas et al., 2002; Schoppek, 1991). For SCS on the other hand, reliability estimates tend to overestimate the true reliability of the measures (Wagener, 2001) as all indicators of performance in SCS are based on the same underlying item structure (see Greiff et al., 2014, for an overview). Correspondingly, the reliability estimates reported for SCS are generally very high ($r_{xx} > .90$; e.g., Kröner et al., 2005; Wagener, 2001). Only MCS tests include multiple, independent items and thus allow for a valid estimation of reliability. These estimates are usually good to very good ($r_{xx} > .80$; Greiff et al., 2013). Repeating our moderation analysis correcting the effect sizes of each type of CPS measure for plausible average reliability coefficients challenged the *low-reliability hypothesis* though. The general pattern of effect sizes remains the same, suggesting that different levels of reliability of the CPS measure used are not causing the divergent findings on the relation between CPS and intelligence.

However, unlike classical measures of CPS or SCS, current MCS tests do not feature some highly complex elements of problem solving such as the recognition and handling of time-delayed effects. Thus, the cognitive demands posed by MCS tests are likely to be relatively closer to those posed by intelligence measures. Following the *different-demands hypothesis* (Rigas & Brehmer, 1999), this might be causing the high correlations of intelligence and CPS scores obtained from MCS tests. In order to test this hypothesis, it would be necessary to develop MCS tests that feature highly complex elements while simultaneously maintaining high levels of reliability.

Finally, our results seem to contradict the *Elshout–Raaheim hypothesis* (Leutner, 2002) that assumes an inverted U-shaped relation between the availability of domain specific knowledge in a CPS task and its correlation with measures of intelligence. Both single complex systems based on LSE and MCS tests are unrelated to any real-world problems and should be equally unaffected by domain specific knowledge. Thus, the *Elshout–Raaheim hypothesis* would predict equally strong effect sizes for the relation between measures based on these two approaches and measures of intelligence. Our finding that MCS tests relate more strongly to intelligence than SCS seems to challenge this hypothesis. In order to test the Elshout-Raaheim hypothesis, CPS measures with comparable psychometric properties but different levels of domain-specific elements would be needed.

8. Implications

Considering the substantial overlap of CPS and intelligence, future research on CPS should focus on theoretically and empirically relating the research conducted on CPS to the vast existing body of research on human abilities.

Next to a more comprehensive theoretical understanding of human abilities, a conflation of research on CPS and more traditional constructs such as intelligence could proof beneficial regarding the advancement of assessment instruments. As the assessment of CPS has predominantly relied on computer-based approaches, CPS researchers went to great effort to maximize the gains from computer-based assessments. Analyzing the behavioral patterns that individuals engage in when dealing with CPS tasks provides insights that go beyond mere final outcome scores and provides access to aspects of the cognitive process underlying specific problem solving behavior. Such in-depth log-file analyses have become technically feasible for CPS research (Scherer, Greiff, & Hautamäki, 2015) and could be extended to intelligence testing (Kröner, 2001) in which the possibilities of computer-based assessment such as log-file data are not fully used yet (Becker, Preckel, Karbach, Raffel, & Spinath, 2015). Getting access to the behaviors displayed and strategies employed by participants in assessments of intelligence could lead to a more thorough understanding of not only the assessment instruments themselves but more importantly of the whole construct of intelligence.

9. Conclusion

In sum, results of the present meta-analysis demonstrate a significant and substantial correlation of CPS and intelligence. Successfully dealing with complex problems requires actively gathering information about a problem in order to later integrate that information to be used to reach a certain goal. The results suggest that a large part of that process involves the cognitive abilities comprising general intelligence. Those with higher intelligence may be better at integrating information

or employ more appropriate strategies in the acquisition of information (Wüstenberg et al., 2014). Thus, we conclude that research on both CPS and intelligence should not only be continued but be symbiotically combined in order to reach a more comprehensive view on human cognitive abilities.

Acknowledgements

We would like to thank Michael McDaniel for his excellent suggestions on the statistical analysis and presentation of our data.

References[2]

Becker, N., Preckel, F., Karbach, J., Raffel, N., & Spinath, F. M. (2015). Die Matrizenkonstruktionsaufgabe [The matrix construction task]. *Diagnostica*, *61*, 22–33. http://dx.doi.org/10.1026/0012-1924/a000111.

Beckmann, J. F. (1994). *Lernen und komplexes Problemlösen. Ein Beitrag zur Konstruktvalidierung von Lerntests [Learning and complex problem-solving. A contribution to the construct validation of tests of learning potential].* Bonn, Germany: Holoshttp://dx.doi.org/10.1080/02724980143000262.

*Beckmann, J., & Guthke, J. (1995). Complex problem solving. Intelligence and learning ability. In P. A. Frensch, & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). http://dx.doi.org/10.4324/9781315806723.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *5*, 1088–1101. http://dx.doi.org/10.2307/2533446.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* Chichester: Wiley.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, *81*, 211–241. http://dx.doi.org/10.1016/0001-6918(92)90019-a.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman and Hall.

*Danner, D. (2011). *Cognitive ability beyond IQ.* Doctoral dissertation University of Heidelberg (Retrieved from http://www.ub.uni-heidelberg.de/archiv/12574).

*Dörner, D., & Kreuzig, H. W. (1983). Problemlösefähigkeit und Intelligenz [Problemsolving ability and intelligence]. *Psychologische Rundschau*, *34*, 185–192. http://dx.doi.org/10.7771/1932-6246.1118.

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (Eds.). (1983). *Lohhausen: Vom Umgang mit Unbestimmtheit und KomplexitätLohausen: On the handling of uncertainty and complexity*. Bern: Huber (Retrieved from http://www.verlag-hanshuber.com/).

Elshout, J. J. (1987). Problem solving and education. In E. DeCorte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction* (pp. 259–273). Oxford: Pergamon (Retrieved from http://ukcatalogue.oup.com/).

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*, 77–96. http://dx.doi.org/10.1207/s15328031us0202_02.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population coefficients vary? *Psychological Methods*, *1*, 444–467. http://dx.doi.org/10.1037/1082-989x.10.4.444.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665–694. http://dx.doi.org/10.1348/000711010x502733.

Frensch, P. A., & Funke, J. (Eds.). (1995). *Complex problem solving: The European perspective*. Hillsdale, NJ: Erlbaum (Retrieved from http://www.psypress.com/).

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, *7*, 69–89. http://dx.doi.org/10.1080/13546780042000046.

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

*Gonzalez, C., Thomas, R. P., & Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. *Intelligence*, *33*, 169–186. http://dx.doi.org/10.1016/j.intell.2004.10.002.

*Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence*, *50*, 100–113. http://dx.doi.org/10.1016/j.intell.2015.02.007.

Greiff, S. (2012). Individualdiagnostik der Problemlösefähigkeit [Diagnostics of problem solving ability on an individual level]. Münster, Germany: Waxmann http://dx.doi.org/10.5539/jedp.v2n1p49

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence*, *41*, 579–596. http://dx.doi.org/10.1016/j.intell.2013.07.012.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2014). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning*, 1–27. http://dx.doi.org/10.1080/13546783.2014.989263 (ahead-of-print).

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

Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs*, *58*, 437–448. http://dx.doi.org/10.1080/03637759109376240.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. http://dx.doi.org/10.3102/10769986006002107.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. http://dx.doi.org/10.1037/1082-989x.3.4.486.

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560. http://dx.doi.org/10.1136/bmj.327.7414.557.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis. Correcting error and bias in research findings* (2nd ed.). Newburypark, CA: Sage Publicationshttp://dx.doi.org/10.1177/1094428106295494.

IBM (2011). *IBM SPSS Statistics for Windows, Version 20.0.0.* Armonk, NY: IBM.

Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. Form 4.* Göttingen, Germany: Hogrefe (Retrieved from http://www.hogrefe.de/).

Joslyn, S., & Hunt, E. (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology: Applied*, *4*, 16–43. http://dx.doi.org/10.1037/1076-898X.4.1.16.

*Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. Rowe (Ed.), *Intelligence: reconceptualization and measurement* (pp. 227–244). Hillsdale, NJ: Lawrence Erlbaum.

*Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation [Assessing intelligence with computer simulations].* Waxmann Verlag.

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

*Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior*, *18*, 685–697. http://dx.doi.org/10.1016/s0747-5632(02)00024-9.

*Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen (Problem solving]. In PISA-Konsortium Deutschland (Ed.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland — Ergebnisse des zweiten internationalen Vergleichs* (pp. 147–175). Münster: Waxmann (Retrieved from http://www.waxmann.com/).

Mayer, H., Hazotte, C., Djaghloul, Y., Latour, T., Sonnleitner, P., Brunner, M., ... Martin, R. (2013). Using complex problem solving simulations for general cognitive ability assessment: The Genetics Lab framework. *International Journal of Information Science and Intelligent System*, *2*, 71–88 (Retrieved from http://www.ijisis.org/).

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. http://dx.doi.org/10.1016/j.intell.2008.08.004.

*Neubert, J. C., Kretzschmar, A., Wüstenberg, S., & Greiff, S. (2014). Extending the assessment of complex problem solving to finite state automata. Embracing heterogeneity. *European Journal of Psychological Assessment*, *31*(3), 181–194. http://dx.doi.org/10.1027/1015-5759/a000224.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*, 354–379. http://dx.doi.org/10.1037/1082-989x.3.3.354.

*Putz-Osterloh, W. (1985). Selbstreflexionen. Testintelligenz und interindividuelle Unterschiede bei der Bewältigung komplexer Probleme [Self-reflections. test intelligence and interindividual differences in solving complex problems]. *Sprache & Kognition*, *4*, 203–216.

R Core Team (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing (http://www.R-project.org/).

Raaheim, K. (1988). Intelligence and task novelty. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence. Vol. 4.* (pp. 73–97). Hillsdale, NJ: Erlbaum (Retrieved from http://www.psypress.com/).

Raven, J. C., Raven, J. C., & De Lemos, M. M. (1958). *Standard progressive matrices.* London: Lewis.

Rigas, G., & Brehmer, B. (1999). Mental processes in intelligence tests and dynamics decision making tasks. In P. Juslin, & H. Montgomerry (Eds.), *Judgment and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 45–65). Hillsdale, NJ: Lawrence Erlbaum.

*Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, *30*, 463–480. http://dx.doi.org/10.1016/s0160-2896(02)00121-6.

*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 assessing complex problem solving. *Psychological Test and Assessment Modeling*, *54*, 54–72.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50. http://dx.doi.org/10.1016/j.intell.2014.10.003.

Schoppek, W. (1991). Spiel und Wirklichkeit — Reliabilität und Validität von Verhaltensmustern in komplexen Situationen [[Game and reality—Reliability and validity of behavior patterns in complex situations)]. *Sprache & Kognition*, *10*, 15–27.

Sternberg, R. J., & Berg, C. A. (1986). Quantitative integration: Definitions of intelligence: A comparison of the 1921 and 1986 symposia. In R. J. Sternberg, & D. K. Detterman (Eds.), *What is intelligence* (pp. 155–162). Norwood, NJ: Ablex.

*Süß, H. M., Kersting, M., & Oberauer, K. (1991). Intelligenz und Wissen als Prädiktoren für Leistungen bei computersimulierten komplexen Problemen [Intelligence and knowledge as predictors of success in computer simulated complex problems]. *Diagnostica*, *37*, 334–352.

Süß, H. M. (1996). *Intelligenz, Wissen und Problemlösen. Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen [Intelligence, knowledge, and*

[2] References marked with an asterisk indicate studies included in the meta-analysis. An asterisk does not precede the in-text citations to studies selected for meta-analysis.

problem solving. Cognitive prerequisites for successful action with computer-simulated problems]. Göttingen: Hogrefe.

Sweeney, L. B., & Sterman, J. D. (2000). Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review*, *16*, 249–286. http://dx.doi.org/10.1002/sdr.198.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research syntheses: Sensitivity analysis using a priory weight functions. *Psychological Methods*, *1*, 428–443. http://dx.doi.org/10.1037/1082-989x.10.4.428.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48 (http://www.jstatsoft.org/v36/i03/).

Viechtbauer, W., & Cheung, M. W. -L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*, 112–125. http://dx.doi.org/10.1002/jrsm.11.

*Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios [Psychological diagnostics using complex scenarios].* Lengerich: Pabst.

Wirth, J., & Klieme, E. (2003). *Computernutzung [Using computers].* Deutsches PISA-Konsortium (Hrsg.), PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland (S. 195–209).

Wittmann, W. W., & Süß, H. M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In R. D. Roberts (Ed.), *Learning and individual differences: Process, trait, and content determinants* (pp. 77–108). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10315-004.

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, *21*, 393–409. http://dx.doi.org/10.1002/sres.653.

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

*Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning*, *19*(1-2), 127–146. http://dx.doi.org/10.1007/s10758-014-9222-8.

4

The logic of success

The relation between complex problem solving skills and university achievement

This article is currently under review:

Stadler, M., Becker, N., Schult, J., Niepel, C., Spinath, F. M., Sparfeldt, J. R., & Greiff, S. (submitted). The logic of success: The relation between complex problem solving skills and university achievement. *Higher Education.*

"The logic of success: The relation between complex problem solving skills and university achievement"

Matthias Stadler[1], Nicolas Becker[2], Johannes Schult[2], Christoph Niepel[1], Frank M. Spinath[2], Jörn R. Sparfeldt[2], & Samuel Greiff[1]

[1] University of Luxembourg, Luxembourg

[2] Saarland University, Germany

Author Note

This research was funded by grants of the Fonds National de la Recherche Luxembourg (ATTRACT "ASKI21"; AFR "COPUS").

The contribution of the third author was supported by the German funds "Bund-Länder-Programm für bessere Studienbedingungen und mehr Qualität in der Lehre ('Qualitätspakt Lehre')" [the joint program of the federal government and states for better study conditions and the quality of teaching in higher education ("the Teaching Quality Pact")] at Saarland University (funding code: 01PL11012). The concept and content of this manuscript was developed by the authors independently of this funding.

Correspondence concerning this article should be addressed to Matthias Stadler, ECCS, University of Luxembourg, Maison des Sciences Humaines, 11, Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg. Phone: +352-466644-5611. E-mail: matthias.stadler@uni.lu

Abstract

The successful completion of a university degree program is accompanied by multiple complex opportunities and challenges, which require students to react accordingly with the skills necessary to meet them. Therefore, the aim of this study was to investigate the role of complex-problem solving (CPS) skills in undergraduate students' university success in two independent samples. In Study 1, 165 university students completed a measure of reasoning as well as a measure of CPS. In addition, students' university GPAs and their subjective evaluation of academic success were collected. CPS made a significant contribution to the explanation of GPAs and the subjective success evaluations even when controlling for reasoning. To further investigate this effect, Study 2 relied on an independent and more heterogeneous sample of 216 university students. The findings of Study 1 were replicated in Study 2. Thus, the results of both studies suggest a link between individual differences in CPS and the abilities necessary to be academically successful in university education.
Keywords: university success; GPA; complex problem solving; intelligence; cognitive ability; structural equation modeling

"The logic of success: The relation between complex problem solving skills and university achievement"

Attending a university is becoming more and more commonplace in modern societies (Pittman & Richmond, 2008), with an increasing number of students enrolling in university programs and societies investing large amounts of money in their educational systems (OECD, 2014). For the individual students, the transition from high school to university life constitutes a critical life event (e.g., Terenzini et al., 1994) with its unique opportunities as well as challenges (Arnett, 2000): opportunities, because the scope of independent exploration of life's possibilities is greater than it will be at any other period of the life course; challenges, because independence and autonomy can also imply disorientation and uncertainty (Arnett, 2000).

Researchers are thus increasingly interested in identifying and examining factors which are related to how students navigate successfully through their university years (Tavernier & Willoughby, 2014). This has resulted in a vast array of cognitive (e.g., intelligence or previous academic achievement; Formazin, Schroeders, Koeller, Wilhelm, & Westmeyer, 2011), noncognitive (e.g., personality traits, motivational factors, self-regulatory learning strategies, students' approaches to learning, or psychosocial contextual influences; for an overview see Richardson, Abraham, & Bond, 2012), and demographic (e.g., age or sociodemographic background; Robbins et al., 2004) factors known to influence students' university success. Within this study we will focus primarily on individual differences in cognitive variables related to students' university performance. Most prominently, previous academic achievement and intelligence have been established as valid predictors of students' grade point average (GPA; Formazin et al., 2011). As an

addition to these established cognitive predictors, academic- related skills such as problem solving were found to be important antecedents of students' success at university (Robbins et al., 2004). In that, individual differences in complex problem-solving skills (CPS), that is, the skills necessary to deal with new and dynamically changing situations (Frensch & Funke, 1995), might provide valuable information in explaining why students succeed differently well at university. Recent research has provided first evidence that CPS is significantly related to academic performance in school (Wüstenberg, Greiff, & Funke, 2012; Greiff et al., 2013) and at university (Stadler, Becker, Greiff, & Spinath, 2015a) with incremental validity over and above intelligence. The present two-study report therefore aims at expanding upon this initial evidence by investigating the relation between university students' skills at dealing with complex problems and their success at university in two studies with independent samples.

Measuring University Success

Understanding university success depends on being able to conceptually define as well as assess it in a reliable and valid way (Richardson et al., 2012). Students' academic performance is usually expressed in terms of GPA representing the mean of the grades received in courses contributing to the final degree (Richardson et al., 2012). GPA is the most widely used and studied measure in tertiary education (Bacon & Bean, 2006; Richardson et al., 2012), is economically available, shows good internal reliability and temporal stability (e.g., Bacon & Bean, 2006), and correlates strongly with variables of interest to educational researchers such as intelligence, motivational strategies, or certain personality traits (Richardson et al., 2012). GPA represents a key criterion for postgraduate selection and employment and has been found to be a valid predictor of socioeconomic success

(Strenze, 2007). Moreover, GPA shows very strong correlations to other indicators of university success such as retention (Robbins et al., 2004). As such, it is an index that is directly meaningful to students, universities, and employers alike and relevant to future training and employment opportunities (Plant, Ericsson, Hill, & Asberg, 2005).

Nonetheless, the use of GPA as an indicator of university success has often been criticized. For example, Johnson (2003) called grade inflation (very good or excellent grades becoming increasingly commonplace) a crisis in university education and argued that every university uses multiple and sometimes very different grading approaches to evaluate students (see also Babcock, 2010). These grading disparities between universities, study programs, and even between different university examiners, as well as the aspect of grade inflation, impair a fair and reliable assessment of students' competencies. This has serious consequences on their future perspectives with respect to completing their university education with a higher GPA and, thus, better career prospects. Thus GPA has, despite its considerable advantages, some noteworthy limitations as a widespread indicator of students' university success.

Beyond a narrow focus on GPA, university success can furthermore be defined as a multidimensional construct with substantial subjective components (Gattiker & Larwood, 1988) such as individual perceptions of accomplishment or future prospects (Aryee, Chay, & Tan, 1994). GPA does not encompass this intrinsic and subjective aspect of success. Furthermore, the notion "university success" respectively "studying successfully" can have many different meanings, such as graduating with a high GPA, graduating as fast as possible, finishing studies and not dropping out earlier on, or the mere subjective satisfaction with the degree (Kunina, Wilhelm, Formazin, Jonkmann, & Schroeders, 2007). In other words, students may, for example, consider a passing grade as either success or failure depending on their

subjective expectations. Correspondingly, researchers have argued that objective and subjective aspects of success should be considered complementary (Duckworth, Weir, Tsukayama, & Kwok, 2012).

Therefore, it is important to assess students' university success based on this multidimensional conceptualization in order to avoid a too narrow coverage of the target construct. In this paper, students' university success will be assessed through their grades as well as through the students' own and subjective evaluation of their university success.

Predicting University Success

Regardless of the specific conception of success, managing a university program requires dealing with a complex system of academic tasks, learning and study behaviors, social obligations, and various other demands that are dynamically changing and whose interrelations are not always obvious (Parker, Summerfeldt, Hogan, & Majeski, 2004). Correspondingly, numerous cognitive (e.g., intelligence or previous academic achievement; Formazin et al., 2011), noncognitive (e.g., personality traits, motivational factors, self-regulatory learning strategies, students' approaches to learning, or psychosocial contextual influences; for an overview see Richardson et al., 2012), and demographic (e.g., age or sociodemographic background; Robbins et al., 2004) factors have been established to influence students' university success.

In this paper, the main focus will be placed on the cognitive predictors of university success. Apart from intelligence, which has been known to be one of the strongest predictors of academic achievement since the early 20th century (e.g., Binet & Simon, 1916; Gottfredson, 2002; Jensen, 1998; Kuncel, Hezlett, & Ones, 2004;

Lubinski, 2004; Roth et al., 2015), other cognitive abilities have been in the focus of researchers recently. Especially in tertiary education, student selection procedures reduce variation in intelligence scores (Furnham, Chamorro-Premuzic, & McDougall, 2003). This is particularly important for universities as highly selective academic institutions (Jensen, 1998). Consequently, factors others than intelligence may add important incremental information to the accurate prediction of performance at the university level.

In addition, substantial differences in the development and prediction of GPA and subjective indicators of university success have been reported (e.g., Harackiewicz, Barron, Tauer, & Elliot, 2002). Whereas cognitive ability consistently predicts university students' GPA, subjective indicators of university success seem to be more closely linked to psychosocial and study skill factors (Robbins, Allen, Casillas, Peterson, & Le, 2006). For instance, Robbins and colleagues (2004) investigated the role of study skill factors as predictors of university outcomes in addition to other well-established cognitive predictors. Their meta-analysis showed academia-related skills, defined as "cognitive, behavioral, and affective tools and abilities necessary to successfully complete task, achieve goals, and manage academic demands" (Robbins et al., 2004, p. 267) to be meaningful predictors of both university GPA and university retention rates with observed mean correlations of $r = .13$ and $r = .30$, respectively.

In addition to these cognitive variables, complex problem solving (CPS) represents a more recently introduced academia-related concept. Recent research has provided initial evidence for the relevance of CPS for academic success at the university (see Stadler, Becker, Gödker, Leutner, & Greiff , 2015b). In this line of research, CPS can be defined as:

(…) the successful interaction with task environments that are dynamic (i.e., change as a function of the user's interventions and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process. (Buchner, cited in Frensch & Funke, 1995, p. 14)

Being able to deal with dynamically changing and partially opaque systems is necessary to be successful at any academic institution. Support for this notion comes from several articles reporting CPS to predict high school grades even beyond measures of general intelligence (Greiff et al., 2013; Wüstenberg et al., 2012; see Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016 for divergent findings) or working memory capacity (Schweizer, Wüstenberg, & Greiff, 2013). Compared to high school, the demands posed by university programs should be even more complex and cognitively challenging. In her model of university success, Ferrett (2000) describes cognitive skills such as time management, preparing for and taking examinations, or using information resources as the focal point of the freshman year experience. In that, university students face a variety of new challenges such as learning and applying study habits in a more complex academic environment and generally discovering how to function as independent and academically successful adults, which requires planning and problem-solving competencies (e.g., acquiring knowledge about new problems or prioritizing subgoals). In other words, students need to solve complex problems. Surprisingly though, only one study has investigated the relation between CPS and university success to date (Stadler et al., 2015a). This study found a substantial relation between CPS and both GPA and subjective university success of business students ($\beta = .38$) that remained significant even after general intelligence was controlled for.

Assessment of CPS

The various elements of academic programs at universities (e.g., courses, teachers, or social obligations) are interrelated in a dynamic system that continues to evolve over time. The skills necessary to deal with such a dynamic system might not be fully captured by static tasks (such as a math problem or an intelligence test item) that do not progress but remain unchanged regardless of the time or the participants' actions (Fischer et al., 2015). To incorporate the dynamic aspect of real-world problem solving, the assessment of CPS has to allow for the problem itself to be dynamic and require the participant to actively interact with the problem in order to understand and manipulate it.

When working on CPS tasks, problem solvers need to manipulate certain input variables of a simulated system (e.g., the duration and intensity of handball training) and observe the resulting changes in a set of outcome variables (e.g., the strength of the players' throws or their endurance). By doing so, problem solvers acquire knowledge (knowledge acquisition phase) about the problem's underlying structure (e.g., high training intensity increases strength but not endurance), which they then apply to reach specific goals (knowledge application phase; Novick & Bassok, 2005). Cognitively, CPS thus involves multiple processes such as causal learning via interaction with the problem (Bühner & Cheng, 2005), hypothesis testing in order to assess the validity of one's own cognitive model (Klahr & Dunbar, 1988), and self-monitoring to avoid inadequate or automatic responses to dynamic changes in the problem (Osman, 2010).

CPS, Intelligence, and Academic Success

When investigating the relation between CPS and university success, it is important to consider the well-established association between measures of CPS and measures of intelligence (e.g., Funke & Frensch, 2007; Wirth & Klieme, 2003; Wüstenberg, et al., 2012). On the one hand, intelligence measures are among the most consistently validated predictors of university success (Richardson et al., 2012), with average correlations of $r = .32$ between intelligence and GPA (corrected for attenuation; Hell, Trapmann, & Schuler, 2007). Intelligence thus possesses a high validity in the prediction of university success and shows incremental validity over high school GPA in predicting university GPA (Formazin et al., 2011). On the other hand, the conceptual and empirical relations of intelligence and CPS need to be considered. CPS and intelligence can theoretically be distinguished by the unique demands complex problems pose. There is, however, considerable theoretical overlap between CPS and intelligence as some characteristic features of CPS such as the integration of information are part of almost every definition of intelligence (Sternberg & Berg, 1986). The majority of studies on the relation between CPS and intelligence correspondingly reports medium to strong correlations between the two constructs (Beckmann & Guthke, 1995; Greiff, Fischer, Stadler, & Wüstenberg, 2015; Greiff et al., 2013). These findings were summarized in a meta-analysis reporting an average correlation of $r = .43$ between CPS and intelligence (Stadler et al., 2015b).

Following this line of thought, it seems necessary to control for the influence of intelligence when investigating the relation between CPS and university success. Otherwise, any associations that are found between CPS and university success might be the result of shared variance of intelligence within the measure of CPS (Wüstenberg et al., 2012).

This Study

Only Stadler and colleagues (2015a) have investigated the relation between CPS and university success to date. While their study provided first evidence supporting the role of CPS in university success, it was severely limited in its generalizability. On the one hand, the sample size used was rather small ($N = 78$) and did not allow for advanced statistical analyses such as structural equation modeling; on the other hand, the sample consisted exclusively of business students and was thus rather homogeneous and limited in terms of generalizability. Finally, the type of CPS measure employed (FSYS; Wagener 2001) was shown to have unsatisfactory reliability (Greiff et al., 2015).

This paper will therefore represent a necessary extension of Stadler and colleague's (2015a) study in various aspects investigating the role of CPS in university success in larger samples, using different indicators of university success for students from various fields of study, and employing more adequate measures of CPS. The aim of Study 1 is to expand upon the results reported by Stadler and colleagues (2015a) using a larger sample thus allowing for latent analyses on the construct level. Study 2 will go even further by investigating the relation between CPS and university success in a very heterogeneous sample. Moreover, considering exam scores as additional criteria and incorporating a longitudinal measurement might reveal further insights regarding the generalizability of these relations.

Study 1

Hypothesis 1: CPS predicts GPA and subjective university success.

Based on the research findings presented above, Study 1 investigated the relation between CPS and students' university success. In line with the findings

reported by Stadler and colleagues (2015a), we expected CPS to significantly predict both GPA and subjective indicators of university success.

Hypothesis 2: CPS predicts GPA and subjective university success even when intelligence is controlled for.

Despite the strong conceptual overlap between CPS and intelligence (e.g., Stadler et al., 2015b; Wüstenberg et al., 2012), the relation between CPS and university success should not be solely due to a shared measurement of intelligence. Thus, we expected to find an incremental validity of CPS in predicting university students' success even when intelligence is controlled for. However, there may be considerable differences in the strength of prediction of the two different constructs. As described above, intelligence is strongly linked to GPA but less strongly to subjective aspects of university success (e.g., Robbins et al., 2006). Correspondingly, CPS should be more important in the prediction of subjective success than GPA after intelligence is controlled for.

Method

Participants. The overall sample consisted of 165 students recruited while attending lectures in the biology ($N = 46$), psychology ($N = 85$), and sports ($N = 34$) departments of a middle-sized German university. Sixty-one percent of the students were female, and the mean age was $M = 22.53$ years ($SD = 3.83$). The majority of students were in their 4th semester of studying at the university (equivalent to the second half of the sophomore year). All students attending the lectures participated in the assessment. Students were told that participation in the study was voluntary and, if they provided an e-mail address, they could receive an individual evaluation of their test results. Participants did not receive any further compensation for their

participation.

Procedure. All tests and questionnaires in Study 1 were conducted solely computer-based. To prevent the uncontrolled influence of different materials on students' performance, the computers were identical in all testing sessions. The entire assessment lasted 90 minutes that is the length of time that students would otherwise have spent in their respective lectures. Before beginning the assessment, students were informed that all personal data would be treated confidentially and would only be used for research purposes. After this, they signed the informed consent sheet approved by the university's data protection agency.

Measures.

Grade point average (GPA). The grade point average (GPA) that had been achieved by participants at the time of the study was employed as a relatively objective measure of university success. GPAs were retrieved from the official university sources (students' had given their approval on the signed consent form). Due to the German grading systems' scoring of 1 representing the best performance grades, for the present analysis, GPAs were reverse-coded so that higher values indicate better performance (with values ranging from 1 to 4), similar to the grading scales used at North American universities.

Subjective university success. Consisting of five items, the scale to measure subjective university success (Stadler et al., 2015a) asked students to rate their agreement with statements such as "I am successful in my studies," "My grades are adequate for my effort," or "My classmates study more successfully than me." Students rated their subjective university success weighed against the amount of effort put in and compared to peers' achievements on a Likert scale ranging from 1 to 5. The

value 1 indicated little and the value 5 indicated great satisfaction with one's own university success. The scale showed good internal consistency ($\alpha = .80$).

Complex problem solving. Individual differences in CPS abilities were assessed using 10 items based on the MicroDYN approach (Greiff, Wüstenberg, & Funke, 2012; Greiff, et al., 2015). This set of items has been shown to provide highly reliable and valid CPS scores (e.g., Greiff et al., 2012). Figure 1 illustrates the type of MicroDYN tasks employed in our study. In the problem depicted here, the test taker is asked to explore the relation between three unspecified training strategies for handball players (labeled A, B, and C) and three outcomes (Motivation, Power of Throw, and Exhaustion). In the course of problem solving, the test taker may systematically vary the use of the three training strategies to determine their effects on the three possible outcomes. It is important to note that, unlike other measures of CPS emulating real-world problems, the underlying relations depicted here are completely arbitrary and do not resemble any real-world setting. Thus, previous knowledge about handball or coaching in general did not provide any advantage for solving the problem. At the end of the knowledge acquisition phase, once knowledge about the system was acquired, participants were asked to plot the assumed relation at the bottom of the task. To reach certain predefined goals in the knowledge application phase (e.g., reach a Motivation value of 20 by adequately adapting the three training methods), the acquired knowledge needed to be applied in a second step.

--- Insert Figure 1 ---

Unlike other measures of CPS, the use of MicroDYN tasks allows for a measurement of individual differences in CPS that is not only theoretically embedded, but also psychometrically confirmed (e.g., Greiff et al., 2012; Greiff et al., 2013). The scoring of students' CPS performance was conducted fully automatized based on

predefined analyses of the results embedded in the testing software. For knowledge acquisition, credit (1 point) was given if the causal model was provided correctly; otherwise, no credit (0 points) was assigned. For knowledge application, credit (1 point) was given if all goals were reached in the application phase; otherwise, no credit was assigned (0 points). The final CPS score was modeled as a second-order factor based on scores of both knowledge acquisition and knowledge application.

Intelligence. In order to determine participants' general intelligence, the well-established Intelligenz-Struktur-Test-Screening (IST-Screening; Liepmann, Beauducel, Brocke, & Nettelnstroth, 2012) was administered. The IST Screening, as a well-established, short (approximately 20 minutes) and economic intelligence measure, consists of the three task groups of verbal analogies, number series, and figural matrices (each consisting of 20 items). The test's publishers report good internal consistencies for all three scales (α = .72 - .90; Liepman et al., 2012). These values were confirmed in our empirical data.

Statistical analysis. To test our hypotheses, we used structural equation modeling (SEM) with weighted least square estimation adjusting means and variances (WLSMV) in Mplus 7.3 (Muthén & Muthén, 1998-2015). Model fit assessment was based on fit indices recommended by Beauducel and Wittmann (2005) and the criteria proposed by Hu and Bentler (1999). For both the Tucker–Lewis index (TLI) and the comparative fit index (CFI), values greater than .90 and .95 were considered to reflect acceptable and good fit to the data, respectively. For the root mean square error of approximation (RMSEA), values of less than .05 and .08 reflect a close fit and a minimally acceptable fit to the data, respectively.

To account for the hierarchical structure of the data due to different university study programs, two dummy variables were created representing the three study

programs with psychology students as the reference group. These dummy variables were added as control variables to all structural models to account for the within-cluster variance in a fixed effects model (Huang, 2016). With a fixed effects model, all variability associated with the cluster level is completely accounted for thereby reducing the problem of omitted variable bias.

Results

Descriptive statistics and measurement models. Table 1 shows the descriptive statistics and observed intercorrelations for all variables included in Study 1. Students' average intelligence scores (M = 47.15; SD = 2.95) were slightly but not significantly [$t(164) = 0.30$; $p = .381$; $d = 0.03$] higher than to be expected based on age and education (norm score = 46.59). However, an ANOVA *comparing the different study programs displayed significant differences in average intelligence scores [$F(2;157) = 5.98$; $p = .003$; $\eta_p^2 = .084$], with psychology students ($M = 48.21$; $SD = 5.28$) averaging significantly higher scores than students studying both biology ($M = 44.57$; $SD = 6.56$) and sports ($M = 43.97$; $SD = 9.03$). Psychology students also scored significantly higher than the corresponding norm sample* [$t(79) = 2.61$; $p = .006$; $d = 0.31$]. The correlations between CPS and both GPA and subjective university success were significant and pointed in the expected direction.

- Insert Table 1 -

Measurement models were established for all latent variables. For subjective university success, all items were defined to load onto one common factor (λ = .45-.92). To limit the parameters to be estimated in the structural models, we aggregated the intelligence items to three parcels for numerical, verbal, and figural content (Little, Cunningham, Shahar, & Widaman, 2002) and had the parcels all load onto one

factor (λ = .51-.72). In line with previous research (e.g., Greiff & Neubert, 2014), CPS was modeled as a higher-order factor consisting of the two latent factors of knowledge acquisition and knowledge application. Both for knowledge acquisition (λ = .50-.99) and knowledge application (λ = .30-.80), all 10 items were defined to load onto one common factor each. As can be seen in Table 2, all measurement models fit very well to the data thus allowing for estimations of the structural models.

- Insert Table 2 -

Structural models. In order to test Hypothesis 1, CPS was specified as a predictor of both subjective university success and GPA. This model represented the data very well as illustrated in the lower part of Table 2. In accordance with Hypothesis 1, CPS significantly predicted subjective university success (β = .32; p = .004; R^2 = .10) and GPA (β = .34; p < .001; R^2 = .12). The correlation between the two criteria subjective university success and GPA was r = .57 (p < .001).

To estimate the incremental validity of CPS over and above intelligence and to avoid issues of multicollinearity resulting from the high latent correlation between CPS and intelligence (r = .81; p < .001), CPS was residualized for intelligence in order to test Hypothesis 2. In this model, intelligence explained 66% (β = .81; p < .001) of the variance in CPS. The remaining residual of CPS, now not sharing any variance with intelligence, as well as intelligence itself were then defined to predict both subjective university success and GPA. In line with Hypothesis 2, the residual of CPS remained a significant predictor of both subjective university success (β = .24; p < .001) and GPA (β = .14; p = .015). As expected, CPS thus predicted subjective success significantly more strongly than GPA after intelligence was controlled for (χ^2 = 2.77; df = 1; p = .048). Intelligence itself also predicted subjective university success (β = .23; p < .001) and GPA (β = .32; p < .001). Intelligence thus predicted

GPA significantly more strongly than subjective university success ($\chi^2 = 8.43$; $df = 1$; $p = .002$). Together, intelligence and CPS explained 12% of the variance in GPA ($R^2 = .12$; $p < .001$) and 11% of the variance in subjective university success ($R^2 = .11$; $p = .007$). The correlation between subjective university success and GPA was $r = .58$ ($p < .001$). As indicated by the fit indices, this model represents the data very well (Table 2) and is illustrated in Figure 2.

--- Figure 2 ---

Discussion of Study 1

The aim of Study 1 was to investigate the relation between CPS and students' university success. In line with our hypotheses and previous research results (Stadler et al., 2015a), CPS was significantly related to both GPA and subjectively appraised success. This relation remained significant and substantial even after intelligence was controlled for.

Regarding the two different indicators of university success, there were considerable differences in the prediction by CPS and intelligence. CPS significantly predicted both GPA and subjective university success. In line with our hypotheses, the relation was significantly stronger between CPS and subjective university success than between CPS and GPA. Intelligence, on the other hand, also predicted both indicators of university success although it was more strongly related to GPA than to subjective university success. This confirms previous findings regarding differential prediction of GPA and subjective indicators of university success (e.g., Robbins et al., 2006) in that GPA is more strongly linked to intelligence while alternative indicators are related to other relevant skills.

In order to estimate the generalizability of these findings, it is necessary to inspect the full correlation matrix. The latent correlation between CPS and

intelligence (r = .81) was slightly higher than suggested by a recent meta-analysis on the relation between CPS and intelligence (Stadler et al., 2015b). This meta-analysis reported a corrected correlation of up to r = .71 (depending on the level of correction for attenuation) for measures of CPS such as the one used in this study. The latent correlation between GPA and intelligence (r = .32) was found to be exactly as was to be expected based on meta-analyses (Hell et al., 2007) reporting average correlations of r = .32.

The relatively homogeneous sample, consisting predominantly of psychology students, suggests a potential range restriction in students' intelligence values, which was confirmed in our data. Psychology students scored significantly higher than students enrolled in the other subject areas as well as the norm sample. This does not come as a surprise considering the German system of selection for students studying in the field of psychology compared to biology or sports. Applicants undergo a competitive selection procedure based on their high school GPA for the limited number of slots available to study psychology at each university. Therefore, this dominance of highly intelligent students within the sample could limit the validity of intelligence (Jensen, 1998). On the other hand, no range restriction (average of .50 with values ranging from 0 to 1) or mean differences [(F(2;157) = 0.16; p = .425; η_p^2 = .002] could be found for CPS values. This finding was expected as the CPS measure was constructed to be used with a university student sample and, accordingly, item difficulties were rather high. A more heterogeneous sample could thus potentially lead to a substantial improvement in the validity of intelligence in the prediction of both subjective university success and GPA.

In summary, the results of Study 1 support the validity of CPS as a predictor of university success, but they raise the question of whether the findings can be

generalized to more heterogeneous samples as well or whether they only hold for specific, highly selective university programs. To further investigate the question of generalizability, we replicated the design of Study 1 in a second study using a more heterogeneous sample.

Study 2

The findings of Study 1 support the hypothesized relevance of CPS as predictor of students' university success. Study 2 expands on Study 1 by replicating the design of Study 1 as closely as possible with the exception of a more heterogeneous sample consisting of students enrolled in diverse study programs. Furthermore, students' scores on a common exam, gathered about three months after the main assessment were added as a longitudinal criterion that was identical for all students. These additions allowed us to investigate the reliability and generalizability of the findings reported in Study 1.

Hypothesis 3: CPS predicts GPA and subjective university success in a heterogeneous sample.

Based on the results found in Study 1, we still expected to find CPS to significantly predict both GPA and subjective indicators of university success.

Hypothesis 4: CPS predicts GPA and subjective university success in a heterogeneous sample even when intelligence is controlled for.

We furthermore expected to find incremental validity of CPS over and above intelligence in predicting university students' GPA and subjective university success. However, the larger variation in university programs should be associated with a larger variation in intelligence. This should increase the validity of intelligence and in turn limit the validity of CPS after controlling for intelligence. This effect should be

particularly strong for GPA and less strong for subjective university success (Robbins et al., 2006).

Hypothesis 5: CPS predicts students' exam results.

Different university study programs have different average grades (Johnson, 2003). Correspondingly, the high heterogeneity in our sample might result in substantial differences in the students' average GPA as well as the frame of reference for the subjective evaluations of their university success. To have a common indicator of university success in addition to students' GPA, students' scores on a course final exam taken by all students participating in Study 2 (see methods section) were additionally included. This additional indicator of university success was gathered several weeks after the main assessment thus allowing for a longitudinal prediction of the students' performance. We expected CPS to predict these final exam scores as well.

Hypothesis 6: CPS predicts students' exam results even when intelligence is controlled for.

This effect should not be solely due to intelligence either, and we expected CPS to show incremental validity in the prediction of students' exam scores over and above intelligence. Similar to GPA, the validity of CPS should be reduced by controlling for intelligence.

Method

Participants. Data from $N = 216$ students (71% women; age: $M = 23.8$ years; $SD = 5.5$ years) in an obligatory introductory lecture on educational assessment for sophomore teacher-education students at a mid-sized German university were used in Study 2. Most students were enrolled in multiple study programs because teachers in

Germany are supposed to teach at least two different school subjects (in addition to the aspects of their study program that are identical for all students striving for teaching degrees, like educational assessment). The most frequent study programs (corresponding with those school subjects to be taught later as a teacher) were German (30%), English (21%), and Mathematics (21%); nevertheless, many other curricula (corresponding to other school subjects) were covered, too.

Procedure. The assessment of demographics and intelligence took place during the first part of the first lecture of the semester. During the following week, the students worked on the CPS tasks online. The final exam covering the entire course took place at the end of the semester. Here, students could choose between two exam dates: either directly at the end of the lecture period (8 weeks after the main assessment) or two months later. The first date was chosen by 58% of the students; 42% of students selected the second date.

Measures. The measures used in Study 1 were also administered in Study 2. CPS was assessed with 10 MicroDYN tasks (Greiff et al., 2012). Intelligence was assessed with the pen-and-paper version of the Intelligenz-Struktur-Test-Screening (IST-Screening; Liepmann et al., 2012). Self-reported university GPAs at the time of the study were reverse-coded so that the best possible passing score was 4 (*very good*) and the worst score was 1 (*sufficient*). The 5-item subjective university success scale ($\alpha = .79$) was completed only by those students who had registered for the second final exam, and it was filled out immediately prior to beginning the final exam. Performance on the course final exam was used as additional indicator of students' university success. Students could, as mentioned, choose freely between the two exam dates. Both exams consisted of 136 multiple-choice items assessing students' competencies and knowledge of educational and psychological assessment; of these,

58 items were identical for both exam dates. A correct response was scored one point and incorrect responses received zero points. Hence, the sum score of the 58 identical items was used as an additional indicator of students' university success.

Statistical analysis. To test our hypotheses, we used structural equation modeling (SEM) with weighted least square estimation adjusting means and variances (WLSMV) in Mplus 7.3 (Muthén & Muthén, 1998-2015) as in Study 1. Again, model fit assessment was based on fit indices recommended by Beauducel and Wittmann (2005) and the criteria proposed by Hu and Bentler (1999) described above.

In line with Study 1, the predictions of GPA and subjective university success were calculated in the same models. The predictions of exam scores on the other hand were calculated in separate models. This was done to consider the time difference between the assessment of GPA and subjective university success and the exam scores, which were gathered several weeks after the main assessment. Whereas the regressions of CPS and intelligence on GPA only represent a statistical relation, the regression of CPS and intelligence on the exam scores and subjective university success represents a real prediction.

Results

Descriptive statistics and measurement models. Table 3 shows the descriptive statistics and intercorrelations for all variables included in Study 2. Students' average intelligence scores ($M = 48.02$; $SD = 5.42$) were slightly higher [$t(249) = 3.79$; $p < .001$; $d = 0.26$] than to be expected based on age and education (norm score = 46.59). The observed correlations between CPS and subjective university success were significant and in the expected direction. The observed correlation between CPS and GPA, however, was not significant.

- Insert Table 3 -

Measurement models were established for all latent variables in the same way as in Study 1. For subjective university success, all items were defined to load onto one common factor (λ = .45-.90). To limit the parameters to be estimated in the structural models, we aggregated the intelligence items to three parcels for numerical, verbal, and figural content (Little et al., 2002) and had the parcels all load onto one factor (λ = .54-.65). CPS was modeled as a higher-order factor consisting of the two latent factors of knowledge acquisition and knowledge application. Both for knowledge acquisition (λ = .51-.98) and knowledge application (λ = .40-.90), all 10 items were defined to load onto one common factor each. All measurement models fit very well to the data thus allowing for estimations of the structural models (see Table 4).

- Insert Table 4 -

Structural models. In order to test Hypothesis 3, we defined the same structural model as for Hypothesis 1. CPS was defined to predict both the latent subjective university success factor and manifest GPA scores. This model represented the data very well as can be seen by the values reported in the second part of Table 4. In accordance with Hypothesis 3, CPS predicted both subjective university success (β = .27; p = .038; R^2 = .07) and GPA (β = .15; p < .001; R^2 = .02). The correlation between subjective university success and GPA was r = .62 (p < .001).

To investigate the incremental validity of CPS (Hypothesis 4), CPS was again residualized for intelligence, which explained 70% (β = .83; p < .001) of the variance in CPS. The remaining residual of CPS, now not sharing any variance with intelligence, as well as intelligence itself were then specified to predict both

subjective university success and GPA. In support of Hypothesis 4, the residual of CPS remained a significant predictor of both subjective university success (β = .17; $p < .001$) and GPA (β = .08; $p < .001$). CPS thus predicted subjective success significantly more strongly than GPA after intelligence was controlled for ($\chi^2 = 3.51$; $df = 1$; $p = .003$). In this model, intelligence itself significantly predicted both subjective university success (β = .41; $p < .001$) and GPA (β = .18; $p < .001$). Combined, intelligence and CPS explained 4% of the variance in GPA ($R^2 = .04$; $p = .001$) and 19% of the variance in subjective university success ($R^2 = .19$; $p = .046$). The correlation between subjective university success and GPA was $r = .63$ ($p < .001$). Thus, this model represented the data very well (Table 4).

Hypothesis 5 stated that CPS predicted students' exam scores. To test this hypothesis, a latent CPS factor indicated by knowledge acquisition and knowledge application was defined to predict manifest exam scores. CPS predicted exam scores significantly (β = .13; $p = .038$; $R^2 = .02$). The fit for this model was very good (Table 4).

To control the effect of CPS in the prediction of students' exam scores for intelligence (Hypothesis 6), CPS was residualized for intelligence (structure identical to Figure 2). In this model, intelligence explained 56% (β = .75; $p < .01$) of the variance in CPS. Contrary to Hypothesis 5, the remaining residual of CPS no longer significantly predicted students' exam scores (β = .06; $p = .295$). Intelligence, on the other hand, predicted students' exam scores significantly (β = .15; $p = .043$). Combined, intelligence and CPS explained significant amounts of variance in students' exam score ($R^2 = .03$; $p = .031$). This model represented the data very well (Table 4).

Discussion of Study 2

The results of Study 2 confirm the role of CPS in the prediction of students' university success. Nonetheless, comparing the results of Study 1 and Study 2 reveals several differences regarding the interplay of CPS and intelligence as predictors of grades and subjective university success. CPS significantly predicted GPA and exam grades as well as students' subjective university success. However, controlling for intelligence resulted in a substantial drop in the relative importance of CPS in the prediction of all three indicators of students' university success. After controlling for intelligence, CPS still predicted GPA and subjective university success but with considerably lower beta weights than found in Study 1. CPS no longer significantly predicted students' exam scores after intelligence was controlled for.

The variations in effect sizes for CPS between the two studies can be interpreted in multiple ways. On the one hand, as already noted, the relatively homogenous sample in Study 1, together with a strong positive selection for cognitive ability, led to a restricted variance in intelligence scores. This restriction in range was less for the sample in Study 2, which also showed a high average intelligence score ($M = 48.02$) but significantly higher variation in intelligence scores [$F(211; 132) = 1.83$; $p < .001$] than the sample in Study 1. This is certainly partially responsible for the relatively low validity of intelligence in the prediction of university GPA in Study 1 and is further supported by the increase in validity of intelligence for the more heterogeneous sample in Study 2. Since we residualized CPS for intelligence in order to estimate the incremental validity of CPS over and above intelligence, a reduced validity of intelligence in Study 1 may have artificially increased the validity coefficients of CPS. It must be noted, however, that our approach of residualizing CPS for intelligence generally benefits the relative

importance of intelligence compared to CPS, as all shared variance is attributed to intelligence (Johnson & LeBreton, 2004).

On the other hand, there also are considerable differences in the structuredness of German university programs (Bargel, Multrus, Ramm, & Bargel, 2009). German programs have a high number of mandatory courses that need to be attended at a certain point of the program. Especially studies in the natural sciences are highly structured leaving little to no room for the students to individualize their studies. Social sciences and humanities programs leave a larger degree of freedom in terms of choosing courses or selecting exam dates (Bargel et al., 2009). This differing number of choices and options might be important, so that students with high CPS are able to use their superior skills to their advantage (Robbins et al., 2004). Given that the sample in Study 1 consisted predominantly of students enrolled in social science programs and the students in Study 2 were enrolled in diverse study programs, this could also have caused the differences in validity for CPS. In a tentative post hoc comparison between students enrolled exclusively in the natural sciences ($N = 55$) and all other students, not exclusively enrolled in natural sciences (see Kaub et al., 2012), the data from Study 2 more strongly supported the validity of CPS in the prediction of students' GPA for students enrolled in social sciences and humanities ($\beta = .21$) than for those enrolled in the natural sciences ($\beta = .06$). This supports the assumption that the validity of CPS in the prediction of university success may rely on the amount of academic freedom students have to individualize their studies. However, due to the post hoc nature and unequal sample sizes, these intriguing results should be interpreted with care and call for additional research.

Taken together, Study 2 replicated most of the findings of Study 1, however, with substantially smaller effect sizes for CPS and substantially larger effect sizes for

intelligence. This supports our interpretation of CPS as a relevant predictor of university success that provides additional information over and above intelligence. The magnitude of this incremental validity may depend on the student population of interest.

General Discussion

Complex problem-solving skills appear to be relevant for the academic achievement of university students. Both studies presented here corroborate the role of this 21st century skill in (tertiary) education with substantial validity. However, the importance of CPS as an additional source of information on students' cognitive ability seems to increase with the selectiveness of university programs (Jensen, 1998). This suggests that individual differences in CPS may be helpful in explaining why highly intelligent students still differ in their academic success. Beyond being smart enough to cope with the academic demands of university, students need to learn to extract relevant information, test hypotheses, and control a dynamically changing environment of interrelated variables (Funke, 2001; Raven, 2000) to succeed in their studies at the university level. In other words, they need to solve complex problems.

Limitations

Nonetheless, some limitations need to be considered in the interpretation of the data. The (mostly) cross-sectional design of both studies calls for caution regarding causal interpretations of the data. Specifically, the correlation between CPS and indicators of university success may represent an increase in CPS as an outcome of university studies rather than individual differences in CPS causing different levels of university success. However, this limitation holds substantially less for the exam scores (Hypothesis 5), which were gathered a considerable amount of time after CPS

was assessed. In line with the interpretation of CPS as a predictor rather than an outcome of university success, CPS predicted these exam scores as well. Notably, the validity coefficients found for intelligence in our studies correspond to those reported in previous longitudinal studies (Hell et al., 2007).

In addition, the choice of operationalization for both CPS and intelligence may have substantially influenced the results. In fact, recent meta-analytic findings have shown that the correlation between CPS and intelligence depends on the operationalizations used (Stadler et al., 2015b). In that, measures of CPS and intelligence (such as those used in the current studies) showed the strongest correlations. Correspondingly, the incremental validity of CPS may have been stronger using different measures of CPS and intelligence (cf. Kretzschmar et al., 2016). However, all measures used in this study are well established and have shown their validity in predicting academic outcomes repeatedly (Liepmann et al., 2012; Wüstenberg et al., 2012), which is not the case for most other measures of CPS (Greiff et al., 2015). Future research may nonetheless investigate whether other operationalizations of CPS lead to a stronger incremental validity of CPS over and above intelligence in the prediction of university success.

Finally, our studies focused exclusively on cognitive predictors of university success. As noted above, various noncognitive predictors of university success have been established as well (for an overview, see Richardson et al., 2012). Little is known about the relation between CPS and noncognitive constructs (Greiff & Neubert, 2014). Wood and Bandura (1989) argue that self-efficacy influences individual use of strategy in CPS tasks and subsequent problem solving attainment whereas Greiff and Neubert (2014) report weak (albeit partly significant) relations between CPS and personality traits. Thus, it stands to reason that future research on

the role of CPS in students' university success needs to include noncognitive factors as well to obtain a more complete picture on the various factors influencing students' success in their studies at the university.

Implications

The major finding of both studies presented here is that individual differences in CPS are related to student's university success and that this difference cannot be reduced to individual differences in intelligence. This confirms previous findings (Stadler et al., 2015b) and provides a solid ground for future research on the role of CPS in university success. CPS tasks may represent a valuable addition to other instruments used in university selection. Besides their validity in predicting relevant outcomes, CPS tasks have been shown to be highly accepted by participants (Sonnleitner et al., 2012), whereas intelligence measures suffer from low acceptance in university selection (Hell & Schuler, 2005). Moreover, computer-based CPS measures provide a vast array of process data (information about single interactions between the problem solver and the task) that may allow may allow researchers a glimpse into the cognitive processes involved in finding successful and unsuccessful solutions to complex problems (Greiff, Wüstenberg, & Avvisati, 2015). Finally, assessing student's individual levels of CPS may help in understanding why some students perform short of their intellectual potential. Helping them to handle the complexity of university studies may thus increase students' success, improve their satisfaction with their education, and ultimately limit the likelihood of a preventable drop-out.

References

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist, 55*, 469-480. doi:10.1037//0003-066X.55.5.469

Aryee, S., Chay, Y. W., & Tan, H. H. (1994). An examination of the antecedents of subjective career success among a managerial sample in Singapore. *Human Relations, 47*, 487-509. doi:10.1177/001872679404700502

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry, 48*, 983-996. doi:10.1111/j.1465-7295.2009.00245.x

Bacon, D. R., & Bean, B. (2006). GPA in research studies: An invaluable but neglected opportunity. *Journal of Marketing Education, 28*, 35-42. doi:10.1177/0273475305284638

Bargel, T., Multrus, F., Ramm, M., & Bargel, H. (2009). *Bachelor-Studierende – Erfahrungen in Studium und Lehre. Eine Zwischenbilanz* [Bachelor students – Experiences in studies and teaching. An interim balance]. Bonn, Berlin, Germany: Bundesministerium für Bildung und Forschung (BMBF).

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75. doi:10.1207/s15328007sem1201_3

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

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kit, Trans.). Baltimore, MD: Williams & Wilkins. doi:10.1037/11069-004

Bühner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 143-168). Cambridge, UK: Cambridge University Press.

Duckworth, A. L., Weir, D., Tsukayama, E., & Kwok, D. (2012). Who does well in life? Conscientious adults excel in both objective and subjective success. *Frontiers in Psychology, 356*, 1-8. doi:10.3389/fpsyg.2012.00356

Ferrett, S. (2000). *Peak performance: Success in college and beyond.* New York: Glencoe/McGraw-Hill.

Fischer, A., Greiff, S., Wüstenberg, S., Fleischer, J., Buchwald, F., & Funke, J. (2015). Assessing analytic and interactive aspects of problem solving competency. *Learning and Individual Differences, 39*, 172-179. doi:10.1016/j.lindif.2015.02.008

Formazin, M., Schroeders, U., Koeller, O., Wilhelm, O., & Westmeyer, H. (2011). Studierendenauswahl im Fach Psychologie. Testentwicklung und Validitätsbefunde [Student selection for psychology. Test development and predictive validity]. *Psychologische Rundschau, 62*, 221-236.

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

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning, 7*, 69-89. doi:10.1080/13546780042000046

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

Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences, 14*, 47-64. doi:10.1016/j.lindif.2003.08.002

Gattiker, U. E., & Larwood, L. (1988). Predictors for managers' career mobility, success, and satisfaction. *Human Relations, 41*, 569-591. doi:10.1177/001872678804100801

Gottfredson, L. S. (2002). g: Highly general and highly practical. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 331-380). Mahwah, NJ: Erlbaum.

Greiff, S., & Neubert, J. C. (2014). On the relation of complex problem solving, personality, fluid intelligence, and academic achievement. *Learning and Individual Differences, 36*, 37-48. doi:10.1016/j.lindif.2014.08.003

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*, 356-382. doi:10.1080/13546783.2014.989263

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence, 50*, 100-113. doi:10.1016/j.intell.2015.02.007

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92-105. doi:10.1016/j.compedu.2015.10.018

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

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond *g*: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*, 364-379. doi:10.1037/a0031856

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94*, 562-575. doi:10.1037/0022-0663.94.3.562.

Hell, B., & Schuler, H. (2005). Verfahren der Studierendenauswahl aus Sicht der Bewerber. [Methods of university student selection from the applicants' view] *Empirische Pädagogik, 19*(4), 361-376.

Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum [A meta-analytic investigation of subject-specific admission tests in the German-speaking countries]. *Empirische Pädagogik, 21*, 251-270.

Hu, L. T., & 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

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education, 84*, 175-196. doi:10.1080/00220973.2014.952397

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

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods, 7*, 238-257. doi:10.1177/1094428104266510

Johnson, V. E. (2003). *Grade inflation: A crisis in college education.* New York, NY: Springer.

Kaub, K., Karbach, J., Biermann, A., Friedrich, A., Bedersdorfer, H.-W., Spinath, F. M., & Brünken, R. (2012). Berufliche Interessensorientierungen und kognitive Leistungsprofile von Lehramtsstudierenden mit unterschiedlichen Fachkombinationen [Vocational interests and cognitive ability of first-year teacher candidates as a function of selected study major] *Zeitschrift für Pädagogische Psychologie, 26*, 233-249. doi:10.1024/1010-0652/a000074

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48. doi:10.1207/s15516709cog1201_1

Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence, 54*, 55-69. doi:10.1016/j.intell.2015.11.004

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148-161. doi:10.1037/0022-3514.86.1.148

Kunina, O., Wilhelm, O., Formazin, M., Jonkmann, K., & Schroeders, U. (2007). Extended criteria and predictors in college admission: Exploring the structure of study success and investigating the validity of domain knowledge. *Psychology Science, 49*, 88-114.

Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). *Intelligenz-Struktur-Test –Screening – IST-Screening* [Intelligence Structure Test - Screening]. Göttingen, Germany: Hogrefe.

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*, 151-173. doi:10.1207/S15328007SEM0902_1

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology, 86*, 96-111. doi:10.1037/0022-3514.86.1.96

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

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

OECD. (2014). *Education at a Glance 2014: OECD Indicators*. OECD Publishing. doi:10.1787/eag-2014-en

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin, 136*, 65-86. doi:10.1037/a0017815

Parker, J. D., Summerfeldt, L. J., Hogan, M. J., & Majeski, S. A. (2004). Emotional intelligence and academic success: Examining the transition from high school to university. *Personality and Individual Differences, 36*, 163-172. doi:10.1016/S0191-8869(03)00076-X

Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology, 30*, 96-116. doi: 10.1016/j.cedpsych.2004.06.001

Pittman, L. D., & Richmond, A. (2008). University belonging, friendship quality, and psychological adjustment during the transition to college. *The Journal of Experimental Education, 76*, 343-362. doi: 10.3200/JEXE.76.4.343-362

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

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353-387. doi:-10.1037/a0026838

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98*, 598 -616. doi:10.1037/0022-0663.98.3.598

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288. doi:10.1037/0033-2909.130.2.261

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. Intelligence, 53, 118–137. doi:10.1016/j.intell.2015.09.002

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences, 24*, 42-52. doi:10.1016/j.lindif.2012.12.011

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*, 54-72.

Stadler, M. J., Becker, N., Greiff, S., & Spinath, F. M. (2015a). The complex route to success: complex problem-solving skills in the prediction of university success. *Higher Education Research & Development,* 1-15. doi:10.1080/07294360.2015.1087387

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015b). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92-101. doi:10.1016/j.intell.2015.09.005

Sternberg, R. J., & Berg, C. A. (1986). Quantitative integration. Definition of intelligence: A comparison of the 1921 and 1986 symposia. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 155-162). Norwood, NJ: Ablex.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*, 401-426. doi:10.1016/j.intell.2006.09.004

Tavernier, R., & Willoughby, T. (2014). Bidirectional associations between sleep (quality and duration) and psychosocial functioning across the university years. *Developmental Psychology*, *50*, 674-682. doi:10.1037/a0034258

Terenzini, P. T., Rendon, L. I., Upcraft, M. L., Millar, S. B., Allison, K. W., Gregg, P. L., & Jalomo, R. (1994). The transition to college: Diverse students, diverse stories. *Research in Higher Education, 35*, 57-73. doi:10.1007/BF02496662

Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios [Psychological diagnostics using complex scenarios]*. Lengerich: Pabst

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy, & Practice, 10*, 329-345. doi:10.1080/0969594032000148172

Wood, R., & Bandura, A. (1989). Impact of conceptions of ability on self-regulatory mechanisms and complex decision making. *Journal of Personality and Social Psychology, 56*, 407-415. doi:10.1037/0022-3514.56.3.407

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

Table 1

Descriptive Statistics and Observed Correlations for Study 1

Variables	*Mean (SD)*	*GPA*	*SUS*	*CPS Acquisition*	*CPS Application*
GPA	*2.31 (0.61)*	-			
SUS	*3.29 (0.59)*	*.33***	-		
CPS Acquisition	*0.52 (0.26)*	*.22***	*.10*	-	
CPS Application	*0.50 (0.23)*	*.19**	*.23***	*.70***	-
Intelligence	*47.15 (2.95)*	*.08*	*.00*	*.24***	*.27***

Note: CPS = complex problem solving; GPA = grade point average; SUS = subjective university success, *$p < .05$, **$p < .01$.

Table 2

Model Fit Indices for Study 1

Model	X^2	*df*	*p*	CFI	TLI	RMSEA
Measurement Models						
SUS	7.51	5	.19	.98	.99	.06
Intelligence	0	0	-	1.00	1.00	.00
CPS	207.81	168	.02	.97	.97	.04
Structural Models						
CPS predicting GPA and SUS	339.63	296	.04	.97	.96	.03
CPS predicting GPA and SUS controlling for Intelligence	409.26	371	.08	.97	.97	.03

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SUS = subjective university success; CPS = complex problem solving; GPA = grade point average.

Table 3

Descriptive Statistics and Correlations for Study 2

Variables	*Mean (SD)*	*GPA*	*SUS*	*Exam scores*	*CPS Acquisition*	*CPS Application*
GPA	*2.18 (0.47)*	-				
SUS	*3.51 (0.65)*	*.52***	-			
Exam scores	*44.75 (6.41)*	*.30***	*.23***	-		
CPS Acquisition	*0.40 (0.27)*	*.04*	*.23***	*.11*	-	
CPS Application	*0.41 (0.24)*	*.00*	*.26***	*.09*	*.71***	-
Intelligence	*48.02 (5.42)*	*.15**	*.46***	*.10*	*.46***	*.43***

Note. CPS = complex problem solving; GPA = grade point average; SUS = subjective university success, *$p < .05$, **$p < .01$.

Table 4

Model Fit Indices for Study 2

Model	X^2	df	*p*	CFI	TLI	RMSEA
Measurement Models						
SUS	3.05	5	.27	1.00	1.00	.00
Intelligence	0	0	–	1.00	1.00	.00
CPS	220.78	170	<.01	.99	.99	.04
Structural Models						
CPS predicting GPA and SUS	375.47	298	<.01	.96	.96	.04
CPS predicting GPA and SUS controlling for intelligence	480.72	375	<.01	.97	.97	.03
CPS predicting exam scores	237.71	189	<.01	.99	.99	.03
CPS predicting exam scores controlling for intelligence	290.78	250	<.01	.99	.99	.02

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SUS = subjective university success; CPS = complex problem solving; GPA = grade point average

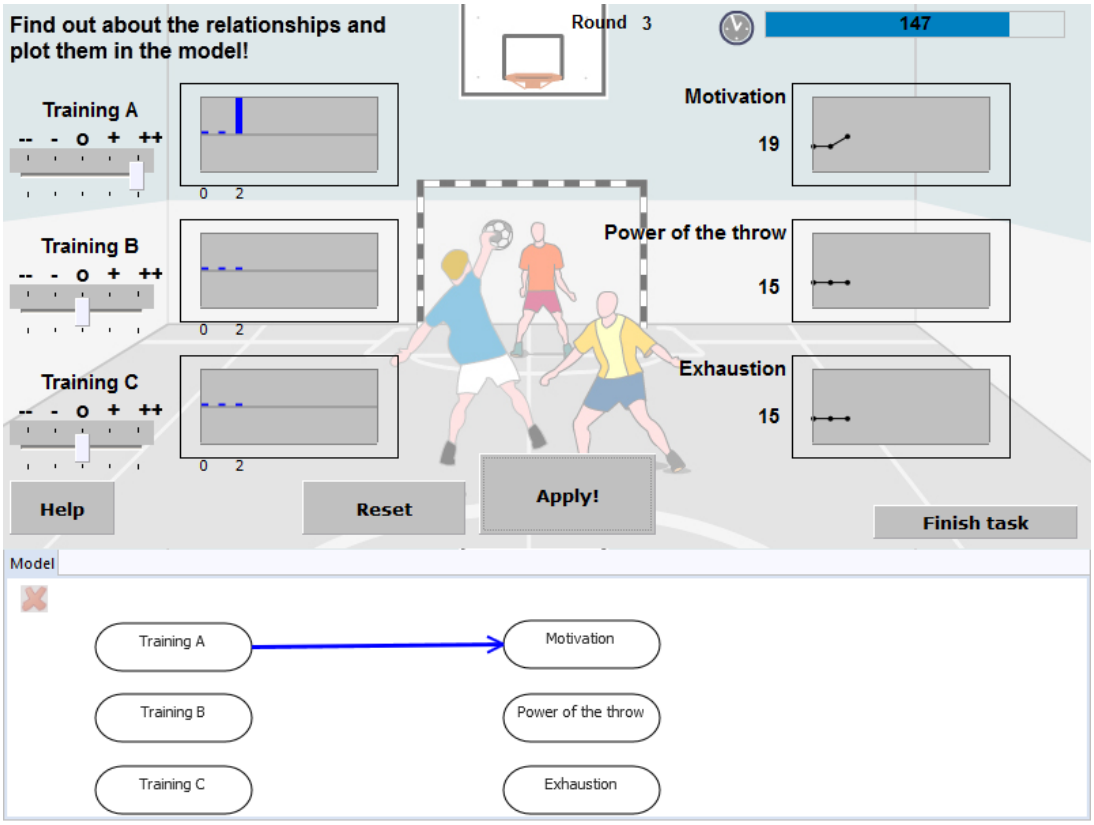

Figure 1. Example of a CPS item based on the MicroDYN approach (Wüstenberg et al., 2012)

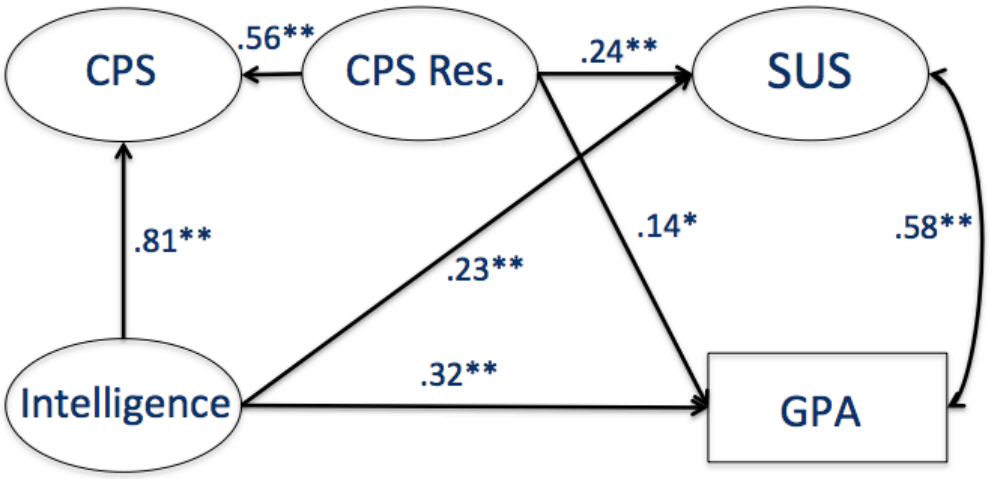

Figure 2. Structural model testing Hypothesis 2 with standardized coefficients. Control variables were omitted for the sake of clarity; CPS = complex problem solving; Res = residual; SUS = subjective university success; GPA = grade point average; *$p < .05$, **$p < .01$.

5

Easily too difficult: Estimating the difficulty of microworlds

This article is available as:

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, *65*, 100-106.

Computers in Human Behavior 65 (2016) 100–106

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

Easily too difficult: Estimating item difficulty in computer simulated microworlds

Matthias Stadler*, Christoph Niepel, Samuel Greiff

University of Luxembourg, Luxembourg

ARTICLE INFO

Article history:
Received 11 June 2016
Received in revised form
5 August 2016
Accepted 15 August 2016

Keywords:
Microworlds
Computer simulation
Problem solving
Linear logistic test model
Item difficulty

ABSTRACT

Dealing with complexity and dynamics is increasingly becoming part of people's everyday lives. The necessity of dealing with complex systems has instigated the use of computer simulations, so-called microworlds (MWs), to assess and study human behavior in complex situations. Although these MWs enjoy great popularity with both practitioners and researchers, their psychometric qualities have been questioned, and studies that have investigated these qualities have been sparse. In particular, only a few studies have investigated the factors that contribute to item difficulty in MWs. To fill this gap, we analyzed data from 3128 Finnish students with a linear logistic test model. Our results suggest that item difficulty in MWs can be almost perfectly predicted by six basic item characteristics, namely, (a) the use and number of eigendynamics, the numbers of (b) input and (c) output variables, the numbers of (d) input and (e) output variables not related to any other variables, and (f) the total number of relations between all variables. In addition, we provide evidence for the necessity of differentiating between the difficulty of controlling an MW (knowledge application) and understanding its underlying structure (knowledge acquisition). Finally, we discuss further theoretical and practical implications of an increased understanding of MWs for their use as assessment instruments.

Running a company, organizing developmental aids for a village in the desert, or coordinating fire fighters during a blaze are highly complex and difficult tasks. Multiple different aspects of the situation need to be considered and the situation changes dynamically. Although we might not face such drastic situations on a daily basis, the world we live in today is becoming more and more complex and dynamic. Just dealing with everyday objects (e.g., phones, computers, automated driving systems) requires us to be aware of their respective connections to other objects or people.

As the complexity of the systems that we interact with in our daily lives grows, so does the importance of research on how we learn to control dynamic environments. Several closely related research areas that focus on how people deal with complex environments have been developed. Most prominent among these research areas are the fields of complex problem solving (CPS; Frensch & Funke, 1995), dynamic decision making (DDM; Brehmer, 1992), systems thinking (Booth-Sweeney & Sterman, 2000), and naturalistic decision making (NDM; Lipshitz, Klein, Orasanu, & Salas, 2001). With the exception of NDM, which focuses primarily on field studies (Klein, 2008), computer simulations and how humans interact with them play integral roles in this research. For example, it would be impossible to have a random participant run an entire company for a short time, but asking the same participant to run a simulated version of the company allows researchers to observe decision making and problem solving in this complex situation. These simulations are supposed to embody the essential characteristics of real-world problems (Gonzalez, Vanyukov, & Martin, 2005), thus representing a compromise between experimental control and realism (Funke, 1992). Throughout this paper, we will use the term *microworlds* (MWs) for reasons of consistency, but several other terms for complex simulations, such as *synthetic task environments* or *high fidelity simulations* have been established as well (for a summary, see Gonzalez et al., 2005).

Despite the considerable use of MWs in both research and practice, many of their relevant characteristics are not yet fully understood, thus limiting their utility. Referring back to the initial examples, it is easy to see how running a company is more difficult than getting used to a new phone, and simulations emulating the former should be harder to understand and control than the latter. But which part of the situation makes one of these tasks harder? Or stated from a psychometrician's point of view, what determines an

* Corresponding author.
E-mail address: matthias.stadler@uni.lu (M. Stadler).

http://dx.doi.org/10.1016/j.chb.2016.08.025
0747-5632/

MW's difficulty, independent of person ability? Whereas the surface differences are obvious, the difficulty of the two MWs should depend on structural characteristics that determine how difficult it is to understand and successfully control an MW. Only a few studies have tried to investigate this question, and rather than conducting an extensive investigation, such studies have focused mostly on individual, specific characteristics of MWs (Kluge, 2008). Therefore, the aim of the current paper is to expand the research on characteristics of MWs that determine their difficulty by systematically analyzing multiple different characteristics. Only by fully understanding the difficulty of MWs can they be optimally fit to specific research questions, samples, and practical requirements.

1. Microworlds in psychological research

The number of complex real-world situations is infinite, and thus it is not surprising that manifold different MWs have been used in psychological research (Funke and Frensch, 2007) ranging from the total control over a city (Dörner, Kreuzig, Reither, & Stäudel, 1983) to working as a fire chief (Brehmer, 1992) or managing a forest (Wagener, 2001). This variety of different MWs is partly due to the initial euphoria over this new test format (Kluge, 2008). MWs were supposed to bridge the gap between field and laboratory research by creating ecologically valid environments that were completely known to and controlled by the researcher (Brehmer & Dörner, 1993). In this, all MWs share some basic characteristics. Gonzales et al., (2005) identified complexity, opaqueness, and dynamics as essential features of MWs. Complexity describes the fact that MWs consist of multiple variables that are related to and thus influenced by each other. These relations between the variables can be expressed by an underlying mathematical structure such as a linear equation and are to a certain degree opaque, meaning that not all of them are always obvious to the person dealing with the MW. Finally, MWs are dynamic, that is, the system's state at time t depends on the state of the system at the previous time t −1 (Rouse, 1981). The term dynamics means that changes in the system can occur either as a result of active manipulations of the system by a participant or through the mere passage of time.

An example of an MW that has been referred to as the "drosophila" of problem solving research (Funke, 2010) and has been used in hundreds of studies is the "Tailorshop" (see Danner, Hagemann, Schankin, Hager, & Funke, 2011). This microworld emulates the workings of a shirt-making company. The system consists of 24 variables that affect each other directly or indirectly (interconnectivity). Of the total of 24 variables, only 21 are visible to a participant who is working on this MW (opaqueness), and only 11 can be manipulated directly, whereas the others change only in response to these manipulations (dynamics). The aim of a problem solver is to maximize the value of the company within a predefined number of steps (i.e., simulated months).

Despite the considerable use of MWs in both research and practice, some defining aspects of MWs such as the Tailorshop are not yet fully understood (Greiff, Wüstenberg, & Funke, 2012). In this paper, we will focus on item difficulty in MWs. In psychological measurement, difficulty is usually defined as a participant's likelihood of responding correctly to an item. However, because MWs rarely have one single correct solution (the problem solvers are relatively free to choose how they will manipulate the system), this definition is not easily applied here (Kluge, 2008). More often, a goal state (e.g., maximizing the total value of the company) that can be achieved through several different courses of action is given. Theoretically, the test developers should be able to specify an optimal or correct solution for achieving the goal state, but given the complexity of many MWs, this is rarely the case (Sager, Barth, Diedam, Engelhart, & Funke, 2011). An exception is the aforementioned Tailorshop for which Sager et al. (2011) attempted to define an optimal solution for every possible state of the system. Even so, various ways of approaching the MW may lead to the same solution. Studies investigating the difficulty of MWs have therefore usually associated an increase in average performance (e.g., a higher company value at the end of the simulation) with decreased difficulty that could be related to differences in the system's underlying structure (e.g., fewer variables that can be directly influenced).

2. Estimating the difficulty of microworlds

In line with this approach, Funke (1983, 1992) was among the first to provide empirical evidence that an MW's difficulty increases with its complexity. In experimental studies, both increasing the number of variables with a fixed number of relations and increasing the number of relations between a fixed number of variables increased an MW's difficulty (see also Greiff, Krkovic, & Nagy, 2014; Kluge, 2004; 2008). A subsequent study investigated the impact of dynamics on the difficulty of MWs (Funke, 1992). In particular, the finding that eigendynamics (i.e., variables affecting themselves) strongly increase the difficulty of MWs has been repeatedly reported (e.g., Funke, 1992; Greiff et al., 2014). A real-world example of eigendynamics can be found in interest rates through which money (or debt) increases over time without additional changes.

Due to the complexity of most MWs, however, all of the previous studies on MWs' difficulty were limited to either specific MWs or rather limited general characteristics such as the number of variables (e.g., Kluge, 2008). Moreover, the great effort related to changing an MW hindered a systematic investigation of the specific characteristics of MWs that influence their difficulty.

An important development toward a more systematic use of MWs in psychological research was suggested in the form of the Multiple Complex Systems approach (MCS; Greiff et al., 2012), which combines multiple small and independent MWs into one test (Greiff, Fischer, Stadler, & Wüstenberg, 2015). The structure of an exemplary MCS microworld is illustrated in Fig. 1. As can be seen, the MW still shows all defining features of an MW, in that there are

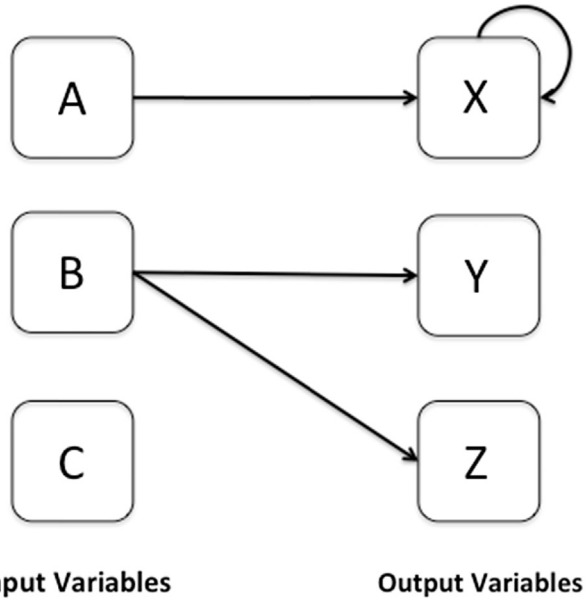

Fig. 1. Abstract example of an MCS microworld based on linear structural equations (adapted from Greiff et al., 2012, p. 192).

different interrelated variables (complexity) only some of which (i.e., the input variables) can be directly manipulated. The relations between the input and output variables are not given to the problem solver and need to be explored by actively interacting with the system (opaqueness). The output variables may change as a result of these interactions or over time (dynamics). In this case, the changes in the system are based on linear structural equations that model the state of the different variables, but there are other MWs that are based on the MCS approach as well (for an overview, see Greiff et al., 2014).

The process of dealing with these different MWs that are based on the MCS approach can be separated into two phases (Novick & Bassok, 2005). First, the problem solver explores the system in order to gain knowledge about the system (knowledge acquisition). In a second step, that knowledge is applied to reach specific target states in the system (knowledge application). Whereas gaining knowledge and applying it are intertwined in real life, separating these processes in an assessment situation allows researchers to obtain more differentiated information about skills, deficits, and possibly the underlying cognitive processes (Greiff et al., 2014).

Tests following the MCS approach thus provide multiple independent scores for knowledge acquisition and knowledge application performance. Given the low complexity of the systems, it is possible for participants to gain complete knowledge of the MW's underlying system, thus allowing for a dichotomous scoring of system knowledge. Similarly, the simplicity of the MWs allows researchers to define achievable control tasks during which certain outcome variables of the system need to reach a certain level. Again, success or failure in this task can be scored dichotomously. The MCS approach thus offers several advantages over the classical approach of using only one large microworld with respect to psychometric properties such as scalability and reliability (for a full review, see Greiff et al., 2013; Greiff et al., 2014) and is well-suited for studies on item difficulty. The independence of the MWs allows for a systematic variation in characteristics, and because the MCS approach provides multiple dichotomous scores of system knowledge and successful system control, this enables the application of complex, IRT-based models of participants' performance. Furthermore, the MCS approach allows the knowledge acquisition phase to be separated from the knowledge application phase. This separation is important because different characteristics may influence the difficulties of these two phases.

Greiff et al. (2014) used these advantages of the MCS approach to apply a linear form of the Rasch Model (RM; Rasch, 1960) called the Linear Logistic Test Model (LLTM; Fischer, 1973) to a number of independent MWs. This model allows researchers to estimate the relative importance of specific characteristics to the difficulty of a set of items (for more details, see below). Greiff et al. (2014) inferred that the number of relations between a varying number of variables as well as the presence of eigendynamics in the MW could account for most of the variance in the MW's difficulty. However, their study investigated only two basic characteristics of MWs (number of relations and eigendynamics) and focused exclusively on the knowledge acquisition phase. In this paper, we extend this paradigm by including different characteristics of MWs to investigate their relative importance for item difficulty in MWs. In addition, we compare the relevance of these characteristics for both knowledge acquisition and knowledge application.

3. The current study

The aim of this paper is to investigate whether the difficulty of a set of MWs constructed within the MCS approach can be described by six essential item characteristics: (a) The use and number of eigendynamics, (b) the number of input variables, (c) the number of output variables, (d) the number of input variables not related to any output variables (i.e., manipulating these variables has no impact on the system and is thus irrelevant for the control of the system), (e) the number of output variables not related to any input variables (i.e., they cannot be controlled and are thus irrelevant for the control of the system), and (f) the total number of relations between all variables. Examples of these characteristics can be seen in Fig. 1 above. In total, this MW has (a) one eigendynamic (Output X), (b) three input variables (Inputs A-C), (c) three output variables (Outputs X-Z), (d) one input variable not related to other variables (Input C), (e) zero output variables not related to other variables, and (f) a total of four relations between all variables. If these six item characteristics completely describe an MW, it should be possible to predict its difficulty with them. This would allow researchers to efficiently create new MWs with predetermined difficulties, thus fitting them optimally to specific populations or research questions.

In the present study, we investigated this hypothesis by going beyond previous work that had concentrated on only a few specific characteristics (e.g., Greiff et al., 2014). Furthermore, it was unclear whether the difficulties of knowledge acquisition and knowledge application would be affected differently by these six item characteristics. By also investigating this distinction, the present study offers a considerably more comprehensive investigation than previous studies in an attempt to further increase our understanding of the determinants of MWs' item difficulty.

4. Method

4.1. Sample and procedure

Our sample consisted of 3128 students attending Grade 6 (N = 1637; 48.7% male; age M = 12.02; SD = 0.41) or Grade 9 (N = 1491; 47.9% male; age M = 14.36; SD = 0.74) in a Finnish municipality. The students were sampled to be representative of the entire population with respect to socioeconomic status and gender. All assessments were administered online with each student working on an individual school computer.[1]

4.2. Instrument

All participants completed a set of nine well-established MWs that followed the MCS approach (Greiff et al., 2014; see above). The MWs consisted of up to three input variables, which were related to up to three output variables. The underlying relations were opaque to the students, and some of the tasks featured eigendynamics. As described above, the students' assignment was to apply adequate strategies to acquire knowledge about the problems' structure (knowledge acquisition) and to apply that knowledge to achieve certain goals (knowledge application). Both the knowledge acquisition and knowledge application phases were scored dichotomously with credit given only if students correctly drew the underlying model or if all goals were reached, respectively. An exemplary MW is illustrated in Fig. 2. In this MW, students are asked to imagine that they are the coach of a handball team and want to find out how different types of training (i.e., Training A, Training B, Training C; left part of Fig. 2) are related to certain

[1] Please note that the performance data employed in this study have been used in previous publications (e.g., Krkovic, Greiff, Kupiainen, Vainikainen, & Hautamäki, 2014; Wüstenberg, Stadler, Hautamäki, & Greiff, 2014; see also Vainikainen, 2014, for more information on the entire assessment battery). However, none of these publications investigated the difficulty of the MWs. Both the research question and every result reported in this study are therefore completely unique to this study.

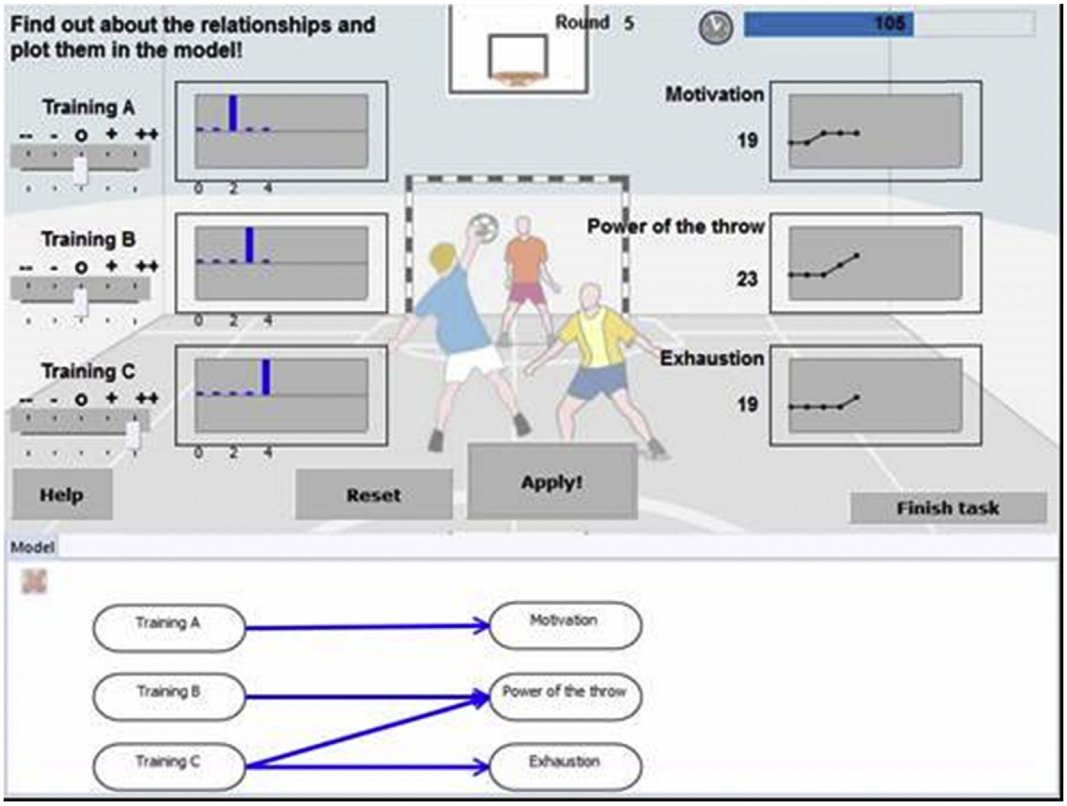

Fig. 2. Screenshot of an exemplary MCS microworld. See text for further details.

characteristics of the team (i.e., Motivation, Power of the throw, Exhaustion; right part of Fig. 2). It is important to note that the relations between the input variables and the output variables are completely arbitrary and not related to any knowledge about handball or coaching in general. Once the knowledge is obtained by systematically varying the input variables (knowledge acquisition phase), the determined relations are plotted in the graph below the task (lower part of Fig. 2). In the second part of the MW (knowledge application phase), predetermined values need to be reached on all outcome variables.

The six item characteristics were distributed across all MWs with no MW including fewer than three or more than five characteristics. The resulting matrix showing the exact distribution of characteristics for each item (design matrix or Q-matrix; Fischer, 1973) is provided in Table 1.

4.3. Data analysis

To estimate the relative importance of the six characteristics for the MWs' difficulty, a linear logistic test model (LLTM; Fischer, 1973) was used. Beginning with the idea that item difficulty can be conceived as a function of certain item characteristics, Fischer developed the model as an elaboration of the more general Rasch Model (RM; Rasch, 1960). The RM states the probability that person j will answer item i correctly on the basis of θ_j, the ability parameter for person j, and σ_i, the difficulty parameter for item i.

$$P(X_{ij}=1|\theta_j,\sigma_i)=\frac{e^{\theta_j-\sigma_i}}{1+e^{\theta_j-\sigma_i}} \quad (1)$$

The LLTM constitutes a linearization of the general Rasch Model (RM; Rasch, 1960). The core assumption is that differences between item difficulties are attributable to item characteristics that vary across the items. What determines an item's difficulty is the number and the nature of the characteristics involved. In the LLTM, the items are scored on these characteristics, and q_{ik} is the score of item i on characteristic k. Estimates from the LLTM include η_k, the weight of k in item difficulty, and θ_j, the ability of person j. The item difficulty σ_i is described as an additive linear function of basic characteristics q_{ik} and the weight of that characteristic η_k:

$$\sigma_i=\sum_{k=1}^{K} q_{ik}\eta_k \quad (2)$$

Replacing σ_i in Equation (1) with Equation (2) yields person j's probability of passing item i in the LLTM:

$$P(X_{ij}=1|\theta_j,q,\eta)=\frac{e^{\theta_j-\Sigma_{k=1}^{K} q_{ik}\eta_k}}{1+e^{\theta_j-\Sigma_{k=1}^{K} q_{ik}\eta_k}} \quad (3)$$

The LLTM includes no error term and therefore assumes that all of the variance in item difficulty can be explained by the basic parameters that have been included (Baghaei & Kubinger, 2015).

In order to estimate the validity of the LLTM, item difficulties (σ) are first determined by applying a general RM followed by an estimation of item difficulties with the LLTM. A high correlation between the two resulting sets of difficulty estimates indicates that the item characteristics provide a good description of the items and thus a good fit of the LLTM (Baghaei & Kubinger, 2015). All analyses were conducted separately for Grades 6 and 9 as well as for knowledge acquisition and knowledge application. All analyses were computed with the R package eRm (Mair, Hatzinger, & Maier, 2012) in R 3.1.1.

Table 1
Design matrix for the 9 MWs and the six characteristics.

	Eigenvalues	Number of input variables	Number of output variables	Irrelevant input variables	Irrelevant output variables	Total number of relations
MW1	0	2	1	0	0	2
MW2	0	2	2	0	0	2
MW3	0	2	2	1	0	2
MW4	0	3	2	0	0	3
MW5	0	3	3	0	0	3
MW6	1	3	2	2	1	2
MW7	0	3	3	0	0	4
MW8	1	3	2	1	0	3
MW9	1	3	3	0	1	4

Note. MW = Microworld.

5. Results

The eta (η) values estimated by the LLTM representing the weight allocated to the specific characteristics in the estimation of item difficulties are displayed in Table 2. Positive eta (η) values indicate a decrease in item difficulty due to the presence of an item characteristic; negative eta (η) values indicate an increase. For knowledge acquisition, all characteristics significantly contributed to the estimation of item difficulty for students in both Grades 6 and 9. The number of eigendynamics was by far the most important characteristic for both age groups (η = −3.32/-3.98), followed by the number of irrelevant input variables (η = −1.77/-1.51) and the total number of relations in the model (η = −1.12/-0.80). All of these characteristics resulted in substantial increases in item difficulty. For the knowledge application phase, on the other hand, the number of irrelevant output variables (η = 1.72/1.94), leading to a decrease in difficulty, and the number of input variables (η = −1.08/-0.93), leading to an increase in difficulty, were most influential. Again, all six item characteristics contributed significantly to the prediction of item difficulty by the LLTM (all *p*s < 0.001). The eta (η) values for students in Grades 6 and 9 were highly correlated in both the knowledge acquisition phase ($r = 0.98$, $p < 0.001$) and the knowledge application phase ($r = 0.94$, $p < 0.001$). This is important as LLTMs with relatively small numbers of items and relatively large numbers of characteristics might overfit the eta (η) values to the data, making it difficult to generalize the results (Fischer, 1973). Finding very similar results in the two independent samples thus provided a cross-validation of the eta (η) values and supported their validity and generalizability to other samples.

Item difficulties (σ) for the RM and the LLTM analyses can be found in Table 3. For both Grades 6 and 9, the LLTM results matched the estimates from the general RM very well. In the knowledge acquisition phase, the correlation between the general RM and the LLTM difficulties approached $r = 1.00$ ($p < 0.001$), suggesting that the six item characteristics almost perfectly described the item difficulties in the MWs. Similarly, the general RM and the LLTM difficulties from the knowledge application phase were very strongly correlated ($r = 0.96$, $p < 0.001$). These results can be taken as evidence that virtually every aspect that was relevant for the difficulty of the MWs was captured by the six characteristics used in this study.

The relation between the RM and LLTM difficulties is further illustrated in Fig. 3 as the relation between the RM and LLTM sigma (σ) values. As can be seen in Fig. 3, the sigma (σ) values of the RM and the LLTM matched each other almost perfectly for the knowledge acquisition phase. For the knowledge application phase, however, the sigma (σ) values that were based on the Rasch Model indicated a lack of variance within the easier items.

Table 2
Eta values for Grade 6 and Grade 9 from the LLTM.

	Knowledge acquisition		Knowledge application	
	Grade 6	Grade 9	Grade 6	Grade 9
Eigenvalues	-3.32^{***}	-3.98^{***}	-0.34^{***}	-1.17^{***}
Number of input variables	0.40^{***}	0.36^{***}	-1.47^{***}	-1.52^{***}
Number of output variables	-0.16^{***}	-0.51^{***}	-1.08^{***}	-0.93^{***}
Irrelevant input variables	-1.77^{***}	-1.51^{***}	0.11^{***}	0.47^{***}
Irrelevant output variables	0.70^{***}	0.83^{***}	1.72^{***}	1.94^{***}
Total number of relations	-1.12^{***}	-0.80^{***}	-0.83^{***}	-0.47^{***}
$r_{(\text{Grade 6, Grade 9})}$	0.98^{***}		0.94^{***}	

$^{***}p < 0.001$.

6. Discussion

The aim of this study was to investigate the validity and relative importance of six essential item characteristics for the prediction of item difficulty in MWs. Our results show that item difficulty in the MWs used in this study could be described almost perfectly as a function of these six item characteristics in an LLTM, a finding that suggests that they cover virtually every aspect relevant for item difficulty. This result replicated previous findings (e.g., Greiff et al., 2014) that had shown that the number of eigendynamics and the total number of relations between all variables were the most important predictors of item difficulty in the knowledge acquisition phase. The current study expanded upon previous studies in showing that various additional item characteristics contributed to the difficulty of MWs as well. Furthermore, we were able to show the necessity of differentiating between the knowledge acquisition and knowledge application phases regarding difficulty because, for knowledge application, the number of irrelevant output variables and the number of input variables had the strongest influences on item difficulty. Thus, different aspects of an MW might determine how difficult it is to acquire new knowledge and to apply this knowledge to reach certain goals. This is important for both the theoretical conception and the empirical use of MWs as the two phases have so far been considered rather equivalent in their psychometric properties (e.g., Wüstenberg, Greiff, & Funke, 2012).

Understanding what constitutes the difficulty of MWs is important for the further use of this item format in psychological assessments (Kluge, 2008). The results of this study suggest that it would be possible to systematically construct both very difficult and rather easy MWs by appropriately combining specific item characteristics. Adding additional input variables, for instance, increases the difficulty of both the knowledge acquisition and knowledge application phases. The addition of irrelevant output variables, on the other hand, is a way to decrease the difficulty of the knowledge application phase while leaving the difficulty of the knowledge acquisition phase relatively unchanged. This is of particular interest for the assessment of giftedness, an area in which many measures fail to systematically reach appropriate

Table 3
Sigma (σ) values based on the general Rasch Model and the LLTM.

	Knowledge acquisition				Knowledge application			
	Rasch Model		LLTM		Rasch Model		LLTM	
	Grade 6	Grade 9	Grade 6	Grade 9	Grade 6	Grade 9	Grade 6	Grade 9
MW1	2.70	3.12	3.23	2.90	3.07	2.87	2.67	2.44
MW2	2.23	1.98	3.08	2.37	1.61	1.57	1.57	1.44
MW3	0.73	0.95	1.36	0.82	1.25	1.21	1.68	1.95
MW4	1.59	1.68	2.38	1.92	−1.68	−1.46	−0.76	−0.71
MW5	1.71	1.72	2.22	1.39	−1.67	−1.57	−1.86	−1.71
MW6	−3.76	−3.84	−2.52	−3.60	1.84	1.85	1.70	1.65
MW7	0.33	0.46	1.13	0.57	−1.62	−1.31	−2.70	−2.22
MW8	−3.44	−3.66	−2.57	−3.72	−1.06	−1.42	−1.00	−1.46
MW9	−2.08	−2.41	−1.42	−2.66	−1.75	−1.74	−1.30	−1.39
$r_{(\beta Rasch,\ \beta LLTM)}$	1.00		1.00		0.96		0.96	

Note. Beta values were standardized to a mean of 0. MW = Microworld; LLTM = Linear logistic test model.

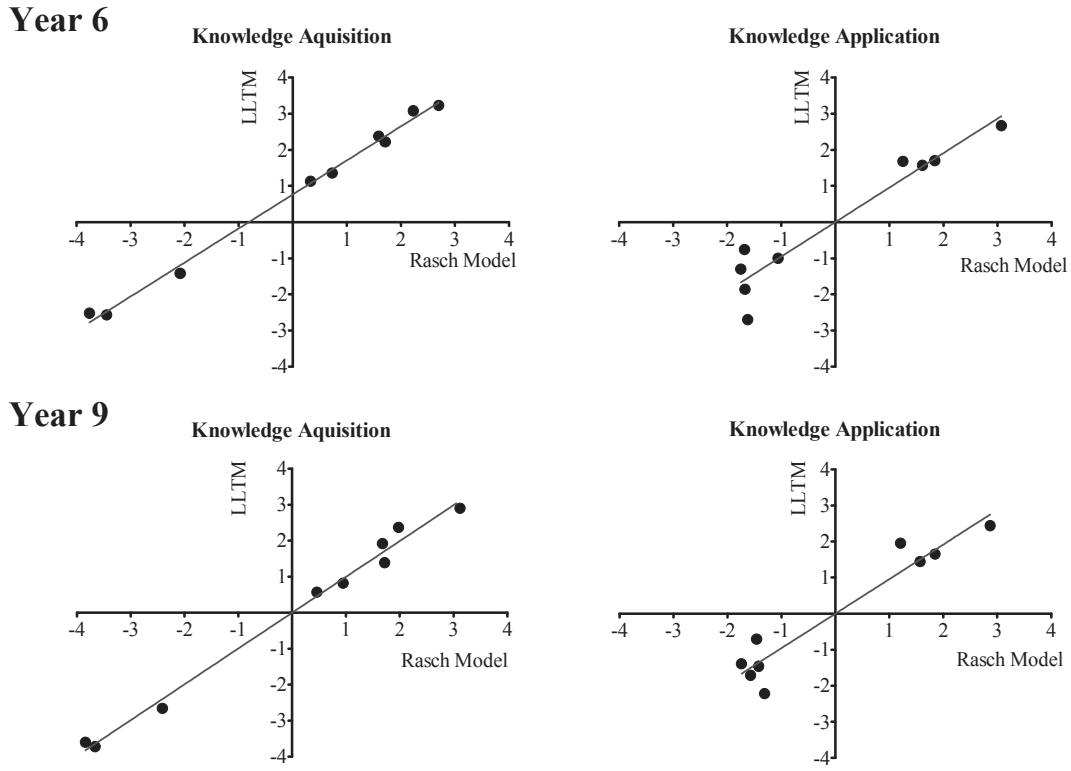

Fig. 3. Graphical representation of the relation between the Rasch Model and the LLTM sigma (σ) values.

levels of item difficulty (e.g., Preckel, 2003), as well as for the assessment of young children, where relatively easy items are required and items that are too difficult may severely harm the children's motivation (e.g., Sonnleitner et al., 2012). Following this line of thought, some of the established microworlds, which consist of up to 2000 variables (Brehmer & Dörner, 1993), should be extremely difficult and hardly solvable to most people. In fact, even the Tailorshop simulation mentioned above consists of a total of 24 variables with over 40 relations between them (Funke, 2010). According to the results presented here, this simulation should be extremely difficult. However, because these very large MWs emulate real-world situations, previous knowledge about the situation may dramatically help in reducing the difficulty. An experienced manager would not need to apply trial-and-error to figure out how an increase in advertising will influence his or her sales. In fact, difficulties that are too high and an overly large influence of previous knowledge were among the major criticisms expressed toward the use of MWs in psychological assessment (e.g., Wittmann & Hattrup, 2004).

The small number of MWs in the LLTM represents a noteworthy limitation of our study. Future studies should expand upon our results by using a larger set of MWs with even more variability in task characteristics. Due to the easily adaptable features of the MCS approach, creating such an item pool is possible. However, our successful cross-validation that was based on the high correlations of eta values for students from Grade 6 and 9 supports the generalizability of our findings.

7. Conclusions

The results of this study provide a way to predetermine the expected difficulty of a microworld when knowledge cannot be

used to gain an advantage, thus providing a good fit to the requirements of the advised assessment and sample. However, future research will need to demonstrate whether the findings presented here can be directly applied to other MWs such as the Tailorshop that are not based on the MCS approach.

Being able to construct new MWs with known item difficulty would also be highly relevant for the use of MWs in high stakes assessments. Having a validated theoretical model of the important characteristics of any item represents an important step toward automatic item generation. Automatic item generation as a means of minimizing the effort necessary to create new items can present a cost efficient and suitable way to use specific item formats, such as MWs, in high stakes testing. Successful implementation of automatic item generation might help reduce the repeated use of single MWs, thus protecting their integrity in high stakes assessments (Arendasy, 2005).

In summary, our findings provide comprehensive information on determinants of MWs' item difficulty that can be used to improve existing assessment instruments, facilitate their use, and instigate future research on this promising item format.

8. Author note

This research was funded by grants from the Fonds National de la Recherche Luxembourg (ATTRACT „ASKI21“ AFR “CoPUS”). Correspondence concerning this article should be addressed to Matthias Stadler, ECCS unit, University of Luxembourg, 11, Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg. Phone: +352-466644-5611. Email: Matthias.stadler@uni.lu. Samuel Greiff is one of two authors of the commercially available COMPRO-test that is based on the multiple complex systems approach and that employs the same assessment principle as MicroDYN. He receives royalty fees for COMPRO.

References

Arendasy, M. (2005). Automatic generation of Rasch-calibrated items: Figural matrices test GEOM and endless-loops test EC. *International Journal of Testing, 5*, 197–224. http://dx.doi.org/10.1207/s15327574ijt0503_2.

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. Practical Assessment. *Research & Evaluation, 20*, 1–11. Retrieved from http://pareonline.net/getvn.asp?v=20&n=1.

Booth-Sweeney, L. B., & Sterman, J. D. (2000). Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review, 16*, 249–286. http://dx.doi.org/10.1002/sdr.198.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta psychologica, 81*, 211–241. http://dx.doi.org/10.1016/0001-6918(92)90019-A.

Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior, 9*, 171–184. http://dx.doi.org/10.1016/0747-5632(93)90005-D.

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*, 323–334. http://dx.doi.org/10.1016/j.intell.2011.06.004.

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität* [*Lohhausen. On dealing with uncertainty and complexity*]. Bern: Huber.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374. http://dx.doi.org/10.1016/0001-6918(73)90003-6.

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

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

Funke, J. (1992). *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung* [*Knowledge about dynamic systems: Acquisition, representation, and application*]. Berlin, Germany: Springer.

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

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

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior, 21*, 273–286. http://dx.doi.org/10.1016/j.chb.2004.02.014.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*, 356–382. http://dx.doi.org/10.1080/13546783.2014.989263.

Greiff, S., Krkovic, K., & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56*, 83–103.

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

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*(2), 364.

Klein, G. (2008). Naturalistic decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*, 456–460. http://dx.doi.org/10.1518/001872008X288385.

Kluge, A. (2004). *Wissenserwerb für das Steuernkomplexer Systeme* [*Knowledge acquisition for controlling complex systems*]. Lengerich, Germany: Pabst Publishers.

Kluge, A. (2008). Performance assessments with microworlds and their difficulty. *Applied Psychological Measurement, 32*, 156–180. http://dx.doi.org/10.1177/0146621607300015.

Krkovic, K., Greiff, S., Kupiainen, S., Vainikainen, M.-P., & Hautamäki, J. (2014). Teacher evaluation of student ability. What roles do teacher gender, student gender, and their interaction play? *Educational Research, 56*, 243–256. http://dx.doi.org/10.1080/00131881.2014.898909.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of behavioral decision making, 14*, 331–352. http://dx.doi.org/10.1002/bdm.381.

Mair, P., Hatzinger, R., & Maier, M. (2012). *eRm: extended Rasch modeling. R package version 0.15-0*. http://CRAN.R-project.org/package=eRm.

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

Preckel, F. (2003). *Diagnostik intellektueller Hochbegabung. Testentwicklung zur Erfassung der fluiden Intelligenz [Assessment of intellectual giftedness]*. Göttingen, Germany: Hogrefe.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rouse, W. B. (1981). Human-computer interaction in the control of dynamic systems. *ACM Computing Surveys (CSUR), 13*, 71–99. http://dx.doi.org/10.1145/356835.356839.

Sager, S., Barth, C. M., Diedam, H., Engelhart, M., & Funke, J. (2011). Optimization as an analysis tool for human complex problem solving. *SIAM Journal on Optimization, 21*(3), 936–959.

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 assessing complex problem solving. *Psychological Test and Assessment Modeling, 54*, 54–72.

Vainikainen, M.-P. (2014). *Finnish primary school pupils' performance in learning to learn assessments: A longitudinal perspective on educational equity*. Helsinki: Picaset: University of Helsinki. Department of Teacher Education Research Reports, 360.

Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios [Psychological diagnostics using complex scenarios]*. Lengerich: Pabst.

Wittmann, W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393–440. http://dx.doi.org/10.1002/sres.653.

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

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge, and Learning, 19*, 127–146. http://dx.doi.org/10.1007/s10758-014-9222-8.

6

General Discussion

6. Discussion

The aim of this thesis was to validate CPS as a construct and to investigate its utility of in the prediction of university success. This research question has never been tackled extensively before and in the four papers that constitute this thesis it could be shown that (1) CPS can be measured reliably and what measurement approach to use best for this thesis, (2) CPS is strongly related but not redundant to intelligence, which supports the idea of CPS as a valuable addition to measures of intelligence in predicting university success, (3) CPS is valid in predicting different indicators of university success and in that shows incremental validity over and above intelligence, and (4) CPS tasks are well understood and can therefore be efficiently created, which is vital for their use in high-stakes assessments such as university applicant selection.

6.1 Implications

CPS tasks thus represent a valuable addition to other instruments used in university applicant selection. As presented in Paper 1, the most suitable CPS tasks for this endeavor are based on the multiple complex systems approach (Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015). Three different sets of tasks following the MCS approach have been developed so far. Of these, the Genetics Lab (Sonnleitner et al., 2012) was specifically designed for young children and may thus not be applicable to university students. The other two sets of tasks – MicroDYN (Greiff, Wüstenberg, & Funke, 2012) and MicroFIN (Neubert, Kretzschmar, Wüstenberg, & Greiff, 2014) – were developed for older students and adults and are therefore appropriate to be used in university selection. The two measures correlate highly and correlate equally strong with intelligence (Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016). The prediction of difficulties described in Paper 4 however, is only possible for MicroDYN tasks. Taken together, this implies that

MicroDYN tasks represent the most adequate measure of CPS to be used in the prediction of university success to date.

The use of MicroDYN tasks (or MCS tasks in general) for the assessment of CPS comes with the cost of high latent correlations between CPS and intelligence as was shown in Paper 2. Latent correlations of $r = .72$ between CPS and intelligence indicate that the additional cognitive demands incorporated in MCS tasks (cf. different demands hypothesis; Rigas & Brehmer, 1999) seem to be limited. Using MicroDYN/ a MCS test as an additional tool in university selection might thus be most adequate when the utility of intelligence is reduced by strong positive selection as would be expected in a highly selective university. Spearman's Law of Diminishing Returns (Jensen, 1998) predicts that the mean correlation among cognitive tests declines as ability level increases. This has implications for the predictive validity of a particular cognitive test. If the mean correlation among cognitive tests is lower for high ability subjects, then the correlation of a particular test with another test will generally be lower for high ability subjects (e.g., Molenaar, Dolan, Wicherts, & van der Maas, 2010). Given that the predictive validity of a test is a test's correlation with a criterion (e.g., GPA), it follows that a test's predictive validity should generally be lower for high ability subjects. Consequently, other factors will add incremental information to the accurate prediction of performance at university level. In other words, the incremental validity of CPS over and above intelligence should be particularly high when the university is very selective and attracts only highly intelligent students.

That CPS does in fact show incremental validity over and above CPS in predicting different indicators of university success was shown in Paper 3. As was to be expected based on previous work (e.g., Robbins et al., 2004; Stadler, Becker,

Greiff, & Spinath, 2015), CPS was particularly valid in predicting subjective university success. This confirms that whereas cognitive ability consistently predicts university students' GPA, subjective indicators of university success seem to be more closely linked to psychosocial and study skill factors (Robbins, Allen, Casillas, Peterson, & Le, 2006). Correspondingly, using CPS to predict university success should be most useful when the focus lies on the subjective aspects of success. This may be relevant to increase university students' satisfaction with their studies or to prevent early drop out (Kunina, Wilhelm, Formazin, Jonkmann, & Schroeders, 2007).

Finally, the results of Paper 4 show that MicroDYN tasks are extremely well understood regarding their defining characteristics. This implies that it is possible to create new tasks with known properties with low effort. They are thus particularly well suited for university applicant selection or other high stake assessment situations. One of the major concerns of high-stakes testing is the integrity of items, which can be severely compromised by repeated use (Way, 2005). Having a working theoretical model of the important characteristics of any item represents an important step toward automatic item generation. Automatic item generation as a means of minimizing the effort necessary to create new items can present a cost efficient and suitable way to use specific item formats in high stakes testing. Successful implementation of automatic item generation might help reduce the repeated use of single tasks, thus protecting their integrity in high stakes assessments (Arendasy, 2005).

In summary, the results of this thesis imply that CPS represents a useful addition to intelligence in understanding and predicting university success.

6.2 Limitations and future research

Some noteworthy limitations calling for further research remain. First, the clear focus on MicroDYN tasks as a measure of CPS may limit the generalizability of

the findings reported. Most importantly, results on the validity of CPS in predicting university success might have been different using different measures of CPS. As outlined in Papers 1 and 2, the MCS approach, which MicroDYN tasks are based on, represents a trade off between qualities and complexity of the tasks (see also Greiff et al., 2015). Using other measures of CPS might therefore have led to different conclusions regarding the predictive power of CPS.

However, despite these limitations the choice of MicroDYN tasks was well justified based on the theoretical considerations and empirical findings reported in Papers 1 and 2. Moreover, Greiff and colleagues (2015) could show that MCS measures are superior to classical measures of CPS in predicting indicators of educational success such as school grades. The results reported in Paper 3 also match those reported by Stadler and colleagues (2015), who did not use a CPS measure based on the MCS approach. Future research should therefore aim to replicate the findings of this study with a broader range of CPS measures in order to achieve a more general operationalization of CPS (see for example Greiff et al., 2013).

The second limitation of this thesis regarding the validity and corresponding utility of CPS in the prediction of university success is the potential trainability of CPS. If deliberate practice can increase the performance on CPS tasks without an actual increase in general CPS competency, this might limit the reliability and thus utility of CPS measures in university applicant selection. Despite more or less explicit recommendations on ways to increase individual CPS competencies and the change of school practices and educational polices in order to foster CPS competence (see OECD, 2014) there is an astonishing lack of empirical research on the trainability of CPS. A noteworthy exception is a study by Kretzschmar and Süß (2016), who report limited transfer effects between different CPS tasks. While this study did not feature

any deliberate practice in the sense of Ericsson (Ericsson, Krampe, & Tesch-Römer, 1993), this work provides first evidences towards a potential, however limited, trainability of general CPS ability. Future research will have to investigate this further in order to estimate how much deliberate training of CPS could diminish the utility of CPS in university selection. Actual increases in individual CPS competency that reflect in improved real-world performance would not influence the validity of CPS in the prediction of university success. Mere increases in CPS task performance without actual increases in CPS competency on the other hand would severely reduce the reliability of these tasks and in result limit their validity.

Related to the issue of trainability, the findings of this thesis (in particular Paper 3) are limited regarding any claims of causality. While it seems intuitive that (1) high CPS, that is the ability to acquire new knowledge about a complex system and use that knowledge to reach specific goals (Frensch & Funke, 1995), will lead to good performance at university, it might also be a result of university training. In other words, the correlation between CPS and indicators of university success may (2) represent an increase in CPS as an outcome of university studies rather than individual differences in CPS causing different levels of university success. In this case, CPS would not be useful as a predictor of university success but rather as a very general criterion of university success indicating how well students are able to deal with complex problems. Finally, there could be (3) a continuous feedback loop between CPS competency and university success with higher individual levels of CPS leading to better performance at university leading in return to even higher levels of CPS.

Due to their cross-sectional nature, the results reported in this thesis cannot rule out any of the three causal relations between CPS and university success. To

approach the question of the causal relation between x and y, a longitudinal study assessing both students' CPS competency and university success multiple times throughout the course of a university program.

Finally, this thesis did not investigate the additional use of CPS process data in the prediction of university success. This could provide additional information about an applicant's skills. CPS testing offers an additional set of completely different information than established predictors of university success usually do. While working on a complex problem, participants can freely explore and interact with the virtual world the problem is set in, allowing the expression of spontaneous and unprompted behavior (Dörner & Wearing, 1995). It is thus possible to gather process data about the way an applicant approaches new problems going beyond mere performance.

To this day, research on the predictive validity of cognitive tasks has mainly focused on measuring final performance, rather than looking at the potential information that could be gained from the process of interacting with a problem (Funke & Frensch, 2007). However, the behaviors displayed in the course of the problem solving process, such as the choice for or against a risky course of action, can be used to deduct non-intellective constructs such as learning strategies (Anderson, 1993), motivational factors (Vollmeyer & Rheinberg, 1999), or personality constructs (Schönbrod & Asendorpf, 2011). For example, it can have great informational value to not actively change a complex system for some time in order to explore its impetus (see Paper 4). Just think of an unfamiliar shower where the water temperature does not change instantly after an adjustment. A constant water temperature can only be reached by waiting until the most recent change has actually become effective. However, since this behavior will not be rewarded immediately, it

takes self-control to employ such a strategy. Deducting traits such as self-control from CPS process data would have the tremendous advantage of them being embedded in the general task of solving the problem at hand so people would not think much about their self-presentation, a problem that questionnaire based research regularly has to face (Hancock & Flowers, 2001).

Future research should therefore investigate in which ways process data of CPS testing can provide additional information that can be used to predict indicators of university success. This dual-purpose of CPS, both measuring intellective skills (overall performance data) and the potential to deducting non-intellective traits (behavioral process data) within one testing session, may make CPS testing a valuable addition to established measures of university applicant selection.

6.3. Conclusion

Taken together, this thesis provided first comprehensive support for the utility of CPS as an additional predictor of university success to be potentially used in university selection. The thesis showed that CPS is related to various indicators of university success and that this relation remains when controlled for the influence of intelligence. Being successful at university thus does not only require being intelligent but also requires being able to understand the complex system that is university life. As this thesis demonstrated, this ability can be measured and differentiated from intelligence.

While not all relevant questions could be answered, the reported results will hopefully provide a solid basis for a large number of future research endeavors. These will be necessary to deal with the new challenges posed by a constantly faster changing world that require people who are able to solve new and highly complex problems. As Albert Einstein put it: "*To raise new questions, new possibilities, to*

regard old problems from a new angle, requires creative imagination and marks real advance in science." (Einstein & Infeld, 1971).

References

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48*(1), 35-44.

Anderson, S. 2003. *The school district role in educational change: A review of the literature.* Ontario: International Centre for Educational Change, Ontario Institute of Studies in Education.

Arendasy, M. (2005). Automatic generation of Rasch-calibrated items: Figural matrices test GEOM and endless-loops test EC. *International Journal of Testing, 5*, 197-224.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry, 48*, 983-996.

Bacon, D. R., & Bean, B. (2006). GPA in research studies: An invaluable but neglected opportunity. *Journal of Marketing Education*, *28*(1), 35-42.

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kit, Trans.). Baltimore, MD: Williams & Wilkins.

Bingham, W. V. (1917). Mentality testing of college students. *Journal of Applied Psychology*, *1*(1), 38-45.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica, 81*, 211–241.

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

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

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

Dörner, D., Kreuzig, H.W., Reither, F., & Stäudel, T. (Eds.). (1983). *Lohhausen: VomUmgang mit Unbestimmtheit und Komplexität* [Lohausen: On the handling of uncertainty and complexity]. Bern: Huber (Retrieved from http://www.verlag-hanshuber.com/).

Duckworth, A. L., Weir, D., Tsukayama, E., & Kwok, D. (2012). Who does well in life? Conscientious adults excel in both objective and subjective success. *Frontiers in Psychology, 356*, 1-8.

Einstein, A., & Infeld, L. (1971). *The evolution of physics: The growth of ideas from early concepts to relativity and quanta*. CUP Archive.

Elias, P., & Purcell, K. (2004). Is mass higher education working? Evidence from the labour market experiences of recent graduates. *National Institute Economic Review*, *190*(1), 60-74.

Elshout, J. J. (1987). Problem solving and education. In E. DeCorte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction* (pp. 259–273). Oxford: Pergamon (Retrieved from http://ukcatalogue.oup.com/).

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review, 100*(3), 363-406.

Ferrett, S. (2000). *Peak performance: Success in college and beyond.* New York: Glencoe/McGraw-Hill.

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

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

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

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

Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences, 14*, 47-64.

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

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence, 41*, 579–596.

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence, 50*, 100–113.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving — A new assessment perspective. *Applied Psychological Measurement, 36*, 189–213.

Hancock, D. R., & Flowers, C. P. (2001). Comparing social desirability responding on World Wide Web and paper-administered surveys. *Educational Technology Research and Development*, *49*(1), 5-13.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94*, 562-575.

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

Johnson, V. E.(2003). *Grade inflation: A crisis in college education.* New York, NY: Springer- Verlag.

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

Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. Rowe (Ed.), *Intelligence: reconceptualization and measurement* (pp. 227–244). Hillsdale, NJ: Lawrence Erlbaum.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology*, *216*(2), 61-73.

Konegen-Greiner, C. (2001). *Studierfähigkeit und Hochschulzugang*. Kölner Texte & Thesen 61. Köln: Deutscher Instituts-Verlag.

Kretzschmar, A., & Süß, H. M. (2015). A study on the training of complex problem solving competence. *Journal of Dynamic Decision Making, 1*(1).

Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence, 54*, 55-69.

Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation* [Assessing intelligence with computer simulations]. Waxmann Verlag.

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

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance. *Psychological bulletin*, *127*(1), 162-181.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148-161.

Kunina, O., Wilhelm, O., Formazin, M., Jonkmann, K., & Schroeders, U. (2007). Extended criteria and predictors in college admission: Exploring the structure of study success and investigating the validity of domain knowledge. *Psychology Science, 49*, 88-114.

Lattner, K. & Haddou, N. (2013). *Abschlussbericht der Studie „Bedingungen von Studienerfolg". Projekt des LearningCenters im Rahmen vom BMBF-Projekt Voneinander Lernen lernen an der Hochschule Osnabrück* (Projektlaufzeit: 10/2012 – 02/2013). Osnabrück.

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

Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen [Problem solving]. In PISA-Konsortium Deutschland (Ed.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland — Ergebnisse des zweiten internationalen Vergleichs* (pp. 147–175). Münster: Waxmann (Retrieved from http://www.waxmann.com/).

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

Mayer, R. E. & Wittrock, M. C. (2006) Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 287-303). Mahwah, NJ: Lawrence Erlbaum.

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

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence, 38*(6), 611-624.

Neubert, J. C., Kretzschmar, A., Wüstenberg, S., & Greiff, S. (2014). Extending the assessment of complex problem solving to finite state automata. Embracing heterogeneity. *European Journal of Psychological Assessment, 31*(3), 181–194.

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

OECD (2014), Education at a Glance 2014: OECD Indicators, OECD Publishing. http://dx.doi.org/10.1787/eag-2014-en

Parker, J. D., Summerfeldt, L. J., Hogan, M. J., & Majeski, S. A. (2004). Emotional intelligence and academic success: Examining the transition from high school to university. *Personality and Individual Differences, 36*, 163-172.

Putz-Osterloh, W. (1985). Selbstreflexionen. Testintelligenz und interindividuelle Unterschiede bei der Bewältigung komplexer Probleme [Self-reflections. test intelligence and interindividual differences in solving complex problems]. *Sprache & Kognition, 4*, 203–216.

Raaheim, K. (1988). Intelligence and task novelty. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*. Vol. 4. (pp. 73–97). Hillsdale, NJ: Erlbaum (Retrieved from http://www.psypress.com/).

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological bulletin*, *138*(2), 353-387.

Rigas, G., & Brehmer, B. (1999). Mental processes in intelligence tests and dynamics decision making tasks. In P. Juslin, & H. Montgomerry (Eds*.), Judgment and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 45–65). Hillsdale, NJ: Lawrence Erlbaum.

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98*, 598 -616.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

Schönbrodt, F. D., & Asendorpf, J. B. (2011). Virtual social environments as a tool for psychological assessment: Dynamics of interaction with a virtual spouse. *Psychological Assessment, 23*(1), 7-14.

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, *24*, 42-52.

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 assessing complex problem solving. *Psychological Test and Assessment Modeling, 54*, 54–72.

Stadler, M. J., Becker, N., Greiff, S., & Spinath, F. M. (2015). The complex route to success: complex problem-solving skills in the prediction of university success. *Higher Education Research & Development,* 1-15.

Statista. (2016). Bewerber und Studienplätze in bundesweiten NC-Studiengängen 2016 | Statistik. Retrieved August 25, 2016, from http://de.statista.com/statistik/daten/studie/36728/umfrage/bewerber-und-studienplaetze-in-bundesweiten-nc-studiengaengen/

Sternberg, R. J., & Berg, C. A. (1986). Quantitative integration: Definitions of intelligence: A comparison of the 1921 and 1986 symposia. In R. J. Sternberg, & D. K. Detterman (Eds.), *What is intelligence* (pp. 155–162). Norwood, NJ: Ablex.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401-426.

Süß, H. M., Kersting, M., & Oberauer, K. (1991). Intelligenz und Wissen als Prädiktoren für Leistungen bei computersimulierten komplexen Problemen [Intelligence and knowledge as predictors of success in computer simulated complex problems]. *Diagnostica, 37*, 334–352.

Vollmeyer, R., & Rheinberg, F. (1999). Motivation and metacognition when learning a complex system. *European Journal of Psychology of Education, 14*(4), 541-554.

Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios* [Psychological diagnostics using complex scenarios]. Lengerich: Pabst.

Walker, I., & Zhu, Y. (2003). Education, earnings and productivity: recent UK evidence. *Labour Market Trends*, *111*(3), 145-152.

Wirth, J., & Klieme, E. (2003). *Computernutzung* [Using computers]. Deutsches PISA Konsortium

Wissenschaftsrat (2004). *Empfehlungen zur Reform des Hochschulzugangs*. (Drucksache 5920/04), Berlin: Wissenschaftsrat.

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

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