The Potential of Complex Problem Solving Scenarios to Assess Students' Cognitive Abilities:

Development, Validation, and Fairness of the Genetics Lab

Dissertation

zur Erlangung des Grades eines Doktors der Philosophie (Dr. phil.)

am Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

vorgelegt von

Mag. rer. nat. Philipp Sonnleitner

Berlin 2015

Erstgutachter: Prof. Dr. Martin Brunner

Zweitgutachter: Prof. Dr. Romain Martin

Tag der Disputation: 20.07.2015

To Ella Madita

Publication list of this cumulative dissertation

Paper 1:

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.

Paper 2:

Sonnleitner, P., Keller, U., Martin, R., Latour, T., & Brunner, M. (in press). Assessing Complex Problem Solving in the Classroom: Meeting Challenges and Opportunities. In B. Csapó & J. Funke (Eds.), *The Nature of Problem Solving*. Paris: OECD.

Paper 3:

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, *41*, 289–305.

Paper 4:

Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, *106*, 681–695.

Abstract

The assessment of students‘ cognitive potential is of central importance to the educational field and is usually done with standardized, paper-and-pencil-based examinations of cognitive ability called intelligence tests (Anastasi & Urbina, 1997; Hunt, 2011; Worthen, White, Fan, & Sudweeks, 1999). Apart from their use for placement decisions–trying to match students with the learning environment which suits them best (Hallinan, 1994; Worthen et al., 1999)–intelligence tests help to ensure optimal cognitive fostering of every child by identifying slow learners and gifted children, and providing a reliable picture of each child's intellectual strengths and weaknesses (Anastasi & Urbina, 1997; Hunt, 2011; Preckel & Baudson, 2013; Shavinina, 2009). More recently, these assessments have been highlighted as a means of evaluating the effects of education itself, as one of the core missions of schooling is to improve students' cognitive abilities by teaching them how to successfully solve problems (Adey et al., 2007; Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Kuhn, 2009; Martinez, 2000, 2013; Mayer & Wittrock, 1996).

Despite their unquestionable successes in fulfilling these purposes, cognitive ability tests have met with serious criticism (Dörner, 1986; Hunt, 2011; Kersting, 1998; Sternberg & Kaufman, 1996). Detractors point to a lack of face validity for predictive capabilities; to the static format of the problems presented, which insufficiently cover all the cognitive abilities that are relevant in real life and do not provide information on test takers' problem representations; and the tests' sensitivity to the emotional state of the subject.

Many scholars have suggested the use of computer-based, complex problem-solving scenarios as a promising alternative to old-style cognitive ability tests (Funke, 2003; Kröner et al., 2005; Rigas, Carling, & Brehmer, 2002). Since such scenarios aim at simulating real-world problems, they are doubtlessly face-valid to test-takers. Computer-based tests also track the subject's problem-solving processes alongside their problem representations (Bennett, Jenkins,

Persky, & Weiss, 2003; Ridgway & McCusker, 2003). Moreover, the possibility of integrating game-like characteristics, such as immediate feedback (Wood, Griffiths, Chappell, & Davies, 2004), may increase motivation and decrease anxiety in the test-takers (McPherson & Burns, 2007; Washburn, 2003). Thus, from a theoretical point of view, complex problem-solving scenarios seem to be attractive candidates to complement – or even replace – traditional, paper-based tests of cognitive abilities.

The present Ph.D. project aims at significantly contributing to the question of whether complex problem-solving scenarios (Dörner, 1986; Funke, 2003) could indeed be used to assess students' cognitive abilities and to determine their potential for this purpose within the educational context. In particular, we have developed a state-of-the art computer-based scenario that assesses complex problem-solving ability called the Genetics Lab (GL) (Sonnleitner, et al., 2012; Sonnleitner, Keller, Martin, Latour, & Brunner, forthcoming) and have examined its construct validity by studying its psychometric structure and relationship to traditional measures of cognitive ability and educational success (Sonnleitner, Keller, Martin, & Brunner, 2013). Finally, to further explore the potential of microworlds, we have investigated whether the Genetics Lab is a fair measure of complex problem-solving ability with respect to students' immigration background (Sonnleitner, Brunner, Keller, & Martin, 2014).

Results showed that the acceptance, and hence the face-validity, was high among 9th graders and that the GL's scores (reflecting the complex problem-solving facets of adequately exploring a problem, gathering knowledge about it, and finally applying this learning to solve the problem) were highly reliable and showed satisfying psychometric characteristics. We were among the first to use a large and representative sample of secondary school students (N = 563) to examine different psychometric conceptualizations of complex problem-solving and their implications for the construct's validity. The results indicate that no matter whether complex

problem solving was modeled as a hierarchical or a faceted construct, it was substantially related to traditional measures of cognitive abilities assessing reasoning and to different indicators of educational success. Controlling for reasoning within a joint hierarchical measurement model, however, revealed that the impressive external validity was largely attributable to the variance that complex problem solving shares with reasoning. This suggests that complex problem solving as a construct has only negligible incremental validity over and above traditional intelligence scales.

Results further showed that the GL is a fair measure of complex problem-solving ability with regard to students' immigration background. Although nonimmigrant students generally outperformed their immigrant peers, such performance differences can largely be explained by differential enrollment in lower academic tracks. Interestingly, the GL's scales were less affected by students' educational background than a traditional paper-pencil-based reasoning scale. Moreover, a fine-grained analysis showed that irrespective of the attended academic track, immigrant students demonstrated a more efficient problem-exploration behavior than their native peers. Taken together, this might point to the potential of computer-based complex problem solving scenarios to identify otherwise hidden cognitive potential in immigrant students.

In sum, the present dissertation gives a differentiated view on the potential of complex problem-solving scenarios for the assessment of students' cognitive abilities within the educational context. Although the strength of these scenarios might not be found in the measurement of something "new" that is not captured by traditional intelligence tests, they provide a novel and innovative approach to measure students' problem-solving processes. This is vital for all educational contexts in which an assessment of students' cognitive potential is needed. The state-of-the art development of the Genetics Lab also points to several implications for modern psychological assessment itself which are discussed at the end of the dissertation.

Keywords: cognitive abilities, intelligence, education, educational assessment, complex problem solving, Genetics Lab (GL), complex problem-solving scenarios, microworlds, incremental validity, external validity, face validity, measurement invariance, computer-based assessment, psychological assessment

Zusammenfassung

Die Erfassung des kognitiven Potenzials von Schülern ist im Bildungssystem von zentraler Bedeutung und wird in der Regel mit standardisierten, Papier-Bleistift basierten Testverfahren, so genannten Intelligenztests, durchgeführt (Anastasi & Urbina, 1997; Hunt, 2011; Worthen, White, Fan, & Sudweeks, 1999). Abgesehen von ihrer Nutzung für Platzierungsentscheidungen, die eine optimale Passung zwischen Schüler und Lernumgebung gewährleisten sollen (Hallinan, 1994; Worthen et al., 1999), erlauben Intelligenztests auch eine adäquate kognitive Förderung von sowohl lernschwachen als auch hochbegabten Schülern, indem sie eine reliable Einschätzung deren intellektueller Stärken und Schwächen ermöglichen (Anastasi & Urbina, 1997; Hunt, 2011; Preckel & Baudson, 2013; Shavinina, 2009). In den letzten Jahren wurde die Bedeutung dieser Testverfahren, vor allem aber auch für die Evaluation von Bildungsprozessen selbst, verstärkt hervorgehoben, stellt die Schulung kognitiver Fähigkeiten durch die Vermittlung von Problemlösestrategien doch einen der zentralen Aufträge des Bildungssystems dar (Adey et al., 2007; Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Kuhn, 2009; Martinez, 2000, 2013; Mayer & Wittrock, 1996).

Trotz ihres unbestrittenen Erfolgs hinsichtlich der Erfüllung der eben genannten Anforderungen, wurden kognitive Fähigkeitstests immer wieder deutlich kritisiert (Dörner, 1986; Hunt, 2011; Kersting, 1998; Sternberg & Kaufman, 1996). Neben dem Fehlen einer offensichtlichen Augenscheinvalidität für ihre prädiktiven Eigenschaften, wurde auch das statische Format der verwendeten Problemtypen kritisiert, das es nur bedingt erlaubt, sämtliche, beim Lösen von Problemen in realen Kontexten aber zentrale, kognitive Fähigkeiten zu erfassen. Zudem bieten klassische Testverfahren keinerlei Informationen hinsichtlich der gewonnenen Problemrepräsentationen der Getesteten und sie erwiesen sich als wenig robust gegenüber emotionalen Einflüssen wie z.B. Testängstlichkeit oder Testmotivation.

In diesem Kontext wurden mehrfach computerbasierte, komplexe Problemlöseszenarien als vielversprechende Alternative zu den traditionellen kognitiven Testverfahren vorgeschlagen (Funke, 2003; Kröner et al., 2005; Rigas, Carling, & Brehmer, 2002). Da derartige Szenarien Probleme aus realen Kontexten simulieren, sind sie zweifelsfrei für die Getesteten augenscheinvalide. Durch die computerbasierte Administration wird es zudem möglich, individuelle Problemlöseprozesse als auch –repräsentationen zu erfassen (Bennett, Jenkins, Persky, & Weiss, 2003; Ridgway & McCusker, 2003). Die zusätzliche Möglichkeit, typische Charakteristiken von Computerspielen zu integrieren, wie bspw. unmittelbares Feedback (Wood, Griffiths, Chappell, & Davies, 2004), könnte sowohl die Motivation bei Getesteten erhöhen als auch eventuelle Testängstlichkeit verringern (McPherson & Burns, 2007; Washburn, 2003). Insofern, zumindest aus theoretischer Sicht, stellen computerbasierte, komplexe Problemlöseszenarien eine interessante Alternative dar, um herkömmliche, Papier-Bleistift basierte, kognitive Fähigkeitstests zu ergänzen bzw. überhaupt zu ersetzen.

Das vorliegende Dissertationsvorhaben hatte zum Ziel, wesentlich zu der Frage beizutragen, inwieweit komplexe Problemlöseszenarien (Dörner, 1986; Funke, 2003) nun tatsächlich zur Erfassung von kognitiven Fähigkeiten bei Schülern eingesetzt werden können und welches Potenzial diese hinsichtlich der oben angesprochenen Anwendungsfelder im Schulsystem haben. Konkret wurde das Genetics Lab (GL) entwickelt, ein State of the Art computerbasiertes Szenario, das es erlaubt, komplexe Problemlösefähigkeit zu erfassen (Sonnleitner, et al., 2012; Sonnleitner, Keller, Martin, Latour, & Brunner, forthcoming). Durch die Analyse der psychometrischen Struktur des GL sowie seines Zusammenhangs mit traditionellen kognitiven Testverfahren und etablierten Maßen von Schulerfolg, konnte zusätzlich seine Konstruktvalidität untersucht werden (Sonnleitner, Keller, Martin, & Brunner, 2013). Um weiters das Potenzial von derartigen Szenarien abzuschätzen, explorierten wir

zusätzlich, inwieweit das GL auch eine faire Einschätzung der komplexen Problemlösefähigkeit in Abhängigkeit vom Migrationshintergrund der Schüler erlaubt (Sonnleitner, Brunner, Keller, & Martin, 2014).

Die Ergebnisse zeigen, dass das GL im Allgemeinen eine hohe Akzeptanz und insofern auch Augenscheinvalidität unter Schülern der 9. Schulstufe genoss und dass die Testkennwerte des GL, welche zentrale Facetten der komplexen Problemlösefähigkeit reflektieren (inwieweit ein Problem adäquat exploriert wird, die gemachten Beobachtungen in deklaratives Wissen übersetzt werden und dieses Wissen zur erfolgreichen Problemlösung angewandt wird), hohe Reliabilität und zufriedenstellende psychometrische Eigenschaften aufwiesen. Als eine der ersten Untersuchungen in diesem Feld bezogen wir uns auf eine große und repräsentative Stichprobe von Sekundarschülern (N = 563), um verschiedene psychometrische Konzeptualisierungen der komplexen Problemlösefähigkeit und deren Implikationen für die Konstruktvalidität zu untersuchen. Die Ergebnisse zeigten, dass komplexes Problemlösen, unabhängig von seiner Modellierung als hierarchisches oder als ein aus einzelnen Facetten bestehendes Konstrukt, stark mit klassischen Maßen kognitiver Fähigkeiten wie Tests zum schlussfolgernden Denken (Reasoning) sowie verschiedenen Indikatoren schulischen Erfolgs assoziiert war. Unter Kontrolle des Einflusses von Reasoning in einem gemeinsamen hierarchischen Messmodel zeigte sich allerdings, dass die beeindruckende externe Validität der Testkennwerte des GL zu einem Großteil auf deren gemeinsame Varianz mit Reasoning zurückgeht. Dies deutet auf eine im Vergleich zu traditionellen Intelligenzskalen eher vernachlässigbare inkrementelle Validität des Konstrukts komplexes Problemlösen hin.

Weitere Analysen zeigten allerdings, dass das GL, unabhängig vom Migrationshintergrund der Schüler, ein faires Maß komplexer Problemlösefähigkeit ist. Obwohl Schüler ohne Migrationshintergrund im Allgemeinen bessere Leistungen zeigten als ihre gleichaltrigen Kollegen, die einen Migrationshintergrund berichteten, konnte dieser

Leistungsvorsprung zum Großteil durch die unterschiedliche Verteilung auf akademische bzw. nicht-akademische Schulzweige erklärt werden. Interessanterweise zeigten sich die Skalen des GL allerdings weniger durch den Schulhintergrund beeinflusst als herkömmliche Papier-Bleistift basierte Reasoning-Skalen. Zudem zeigte eine detaillierte Analyse, dass, unabhängig vom besuchten Schulzweig, Schüler mit Migrationshintergrund eine deutlich effizientere Strategie anwendeten um Probleme zu explorieren als ihre Mitschüler. Zusammengefasst deuten die Befunde auf ein eventuelles Potenzial von computerbasierten, komplexen Problemlöseszenarien hin, anderweitig schwer erfassbares, also verstecktes kognitives Potenzial bei Schülern mit Migrationshintergrund messbar zu machen.

In Summe eröffnet das vorliegende Dissertationsprojekt einen differenzierten Blick auf das Potenzial von komplexen Problemlöseszenarien für die Erfassung kognitiver Fähigkeiten im schulischen Kontext. Obwohl die Stärken dieser Szenarien wohl eher nicht in der Messung von etwas „Neuem“ liegen, das nicht schon durch bereits bestehende Intelligenztests erfasst wird, bieten sie einen neuartigen und innovativen Ansatz, um individuelle Problemlöseprozesse bei Schülern zu messen. Dies ist allerdings zentral für sämtliche schulischen Bereiche, in denen die Abschätzung des kognitiven Potenzials von Schülern notwendig ist. Die State of the Art Entwicklung des Genetics Labs beinhaltet zusätzlich zahlreiche Implikationen für moderne psychologische Diagnostik, welche am Ende der Dissertation diskutiert werden.

Schlüsselbegriffe: Kognitive Fähigkeiten, Kognitives Potenzial, Intelligenz, Komplexes Problemlösen, Genetics Lab (GL), Komplexe Problemlöseszenarien, Microworlds, Inkrementelle Validität, Externe Validität, Augenscheinvalidität, Messinvarianz, Computerbasiertes Testen, Computerbasierte Diagnostik, Prozessmaße, Schuleignungsdiagnostik, Pädagogische Diagnostik, Psychologische Diagnostik

Declaration

I hereby declare that the present dissertation is my own work and that, to the best of my knowledge and belief, it contains no materials previously published or written by another person except where due acknowledgement has been made in the text. I furthermore declare that this work has not been submitted for the award of any other degree or diploma at a university other than the Free University of Berlin.

Philipp Sonnleitner

4th June 2015

Contents

Chapter I –Theoretical background

1.1. The need to assess students' cognitive abilities

Children greatly differ in the ways that they are able to think. Some children think faster than others (Kail & Salthouse, 1994), and some can think of more things at the same time than others (Alloway, Gathercole, Willis, & Adams, 2004; Hornung, Brunner, Reuter, & Martin, 2011). While some struggle to manage multiple tasks at a time, others of their peers have no difficulty at all (Schaefer, Krampe, Lindenberger, & Baltes, 2008). These differences can be attributed in part to the varying stages of cognitive development (Cattell, 1971; Kail & Salthouse, 1994; Piaget, 1969), but also to stable inter-individual differences (Deary, Whalley, Lemmon, Crawford, & Starr, 2000; Schaie, 1996; Schalke et al., 2013). Such cognitive abilities, however, are the cornerstone of knowledge acquisition, learning, and problem solving, and differences in them determine to a large degree a child's success as a student in school (Alloway & Alloway, 2010; Engel de Abreu, Gathercole, & Martin, 2011; Hornung, Schiltz, Brunner, & Martin, 2014). Consequently, school systems that are aimed at providing optimal fostering for every child face several challenges in the light of these cognitive differences (cf. Anastasi & Urbina, 1997; Worthen, White, Fan, & Sudweeks, 1999).

As will be outlined, solutions to these challenges were mostly based on a reliable assessment of students' cognitive abilities, which subsequently served as a basis for further decisions (Anastasi & Urbina, 1997; Hunt, 2011; Worthen et al., 1999). The present dissertation addresses the question of whether the traditionally-used measuring instruments (i.e., intelligence tests) are still adequate or if they should be complemented – or even replaced

by – innovative, computer-based assessment tools (i.e., complex problem-solving scenarios) that offer new insights into the way students think and approach problems.

1.1.1. Placement decisions – Predicting future performance

What makes teaching a common curriculum both interesting and challenging for every teacher is that "given any task, students progress at different rates and achieve different levels of mastery" (Worthen et al., 1999, p.474). A common solution to this cognitive heterogeneity has traditionally been to place students in ability-dependent school tracks, or instructional groups that should provide a learning environment suitable for the individual's capabilities (Hallinan, 1994; Worthen et al., 1999). Although it is hotly debated and has been the subject of numerous critiques (e.g. Hattie, 2009; Oakes, 2005), educational tracking is still common practice, particularly in German-speaking countries (i.e., Austria, Germany, Luxembourg, Switzerland), the Netherlands, and – in a weaker form – the United States.

In this sort of educational tracking, students are typically assigned to a specific secondary-school track. The decision of which track best fits the student is primarily based on their previous grades, the preferences of the student's parents, the student's main elementary school teacher's recommendation, or a combination of several criteria (Baeriswyl, Wandeler, Trautwein, & Oswald, 2006; Ditton, Krüsken, & Schauenberg, 2005; Klapproth, Glock, Böhmer, Krolak-Schwerdt, & Martin, 2012; Roeder, 1997). It has been shown, however, that this decision and most of the criteria themselves are heavily influenced by several other, irrelevant factors, such as the socioeconomic status and the educational attainment of the student's parents and the student's gender and immigration background, clearly reducing the decision's prognostic validity (Deißner, 2013; Ditton et al., 2005; Klapproth et al., 2012; Roeder, 1997; Schnabel & Schwippert, 2000; Tiedemann & Billmann-Mahecha, 2007). Thus,

it has been argued that "pure," language-reduced measures of students' cognitive abilities be used to aid these tracking decisions, since such indicators have been proven to be objective and highly predictive of later educational performance (cf. Heller, 1991; Maaz, Baeriswyl, & Trautwein, 2013).

1.1.2. Too "slow" or too "fast"? - Identifying cognitive extremes

A second challenge lies in dealing with students at the extremes of the cognitive ability distribution who do not necessarily fit into regular tracks (e.g., Anastasi & Urbina, 1997; Worthen et al., 1999). This problem was evident from the very beginning of compulsory education. For example, in 1904, French teachers increasingly complained about "slow" children that were not able to follow the standard curriculum and thus constrained their peers in learning (cf. Hunt, 2011; Nicolas, Andrieu, Croizet, Sanitioso, & Burman, 2013). On the other hand, teachers also noticed children that were more cognitively advanced than their peers and thus remained insufficiently challenged and bored, and worked below their full potential throughout the curriculum (Preckel & Baudson, 2013; Shavinina, 2009; Worthen et al., 1999). It became clear very quickly that both groups – children with learning disabilities and gifted children – would need special treatment that a "one-size-fits-all" school system oriented toward students with average cognitive abilities cannot provide.

The first attempts to deal with this issue were mainly economically motivated and simply strived for (a) weeding out students that slowed down the curriculum or (b) optimally fostering the "human resource" of gifted students for society. Although this perspective is still present, especially in the giftedness debate, over the years a more humanistic perspective has prevailed, one which acknowledges the special needs of both groups and focuses on their right to remediation or an optimal educational treatment (e.g., Grigorenko, 2015; Hunt, 2011;

Mcclain & Pfeiffer, 2012; Nicolas et al., 2013; Preckel & Baudson, 2013; Shavinina, 2009). Central to all solutions to this challenge, however, is the correct identification of learning disabilities or giftedness by diagnosing students' cognitive abilities at a very early age.

1.1.3. Where to go? – Identifying strengths and weaknesses

At a certain point of the school career, usually when compulsory education ends, each student has to decide whether to make the transition from school to the workplace or to stay in school and perhaps pursue an academic career. Schools are often obliged to aid this decision by providing individual counseling. Comparable to tracking decisions (see 1.1.1.), such career choices are influenced by a variety of factors. Besides personality and interests, students are also influenced in their decision by their gender, their ethnic background, their socio-economic status, and the occupations of their parents and peers (e.g., Dick & Rallis, 1991; Tang, Fouad, & Smith, 1999). Compared to the inappropriateness of many of these factors to predict a student's suitability for a job, however, an impressive amount of research has accumulated which clearly shows the high importance of students' cognitive abilities in this context (cf. Gottfredson, 1997; Hunt, 2011; Schmidt & Hunter, 1998).

Hence, not considering a student's cognitive potential when reflecting on his or her professional future may lead to less than ideal career choices. Thus, schools building on a reliable identification of individual (cognitive) strengths and weaknesses when counseling students not only increase the likelihood of an informed decision on the best-fitting vocational or educational environment, but also help to avoid suboptimal career choices that are due to the student's family background (cf. Anastasi & Urbina, 1997; Heller, 1991; Worthen et al., 1999).

1.1.4. Raw material or product? - Evaluating the success of education

When it comes to the relationship between students' cognitive abilities and their education, the prevailing view among educators rests upon the assumption that students bring their cognitive abilities to school and, depending on those cognitive abilities, the students are more or less successful. In other words, students' cognitive abilities are the "raw material" that education transforms into competencies like reading, numeracy, and knowledge (Adey, Csapó, Demetriou, Hautamäki, & Shayer, 2007; Martinez, 2000, 2013). This assumption is also somewhat reflected in the idea of school tracking (see 1.1.1.), which sorts students according to their (initial) cognitive abilities into differentially demanding curricula. At the same time, it is the ultimate goal of education to prepare students to function well in our society (cf. Brock & Alexiadou, 2013) and, given the undisputed importance of cognitive abilities in this context (see above, Gottfredson, 1997; Hunt, 2011), it seems odd that their malleability – or even their training – is not on the agenda.

Meanwhile, an impressive body of research has accumulated demonstrating that cognitive abilities are indeed plastic and to a certain extent trainable (cf. Adey et al., 2007; Hunt, 2011; Hunt & Jaeggi, 2013; Martinez, 2000, 2013). In addition, it has been unanimously shown that education itself is already improving these cognitive abilities (Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Gustafsson, 2008), leading to the somewhat paradoxical conclusion that education implicitly trains cognitive abilities, without explicitly reflecting upon it. Crucially though, ignoring research on cognitive abilities may sometimes cause suboptimal policy making in this domain (cf. Brunner, 2008). Reasons for this neglect are manifold (see also section 1.2.3.), and may in large parts be grounded in the (erroneous) idea of immutable and fixed cognitive abilities – rendering the teacher's job somewhat hopeless. But there are also mixed results concerning the best way of training such abilities, and thus no firm guidance

concerning the specific design of the training (Adey et al., 2007; Hunt, 2011; Hunt & Jaeggi, 2013).

During the last few years, however, there has been a recurring demand to explicitly teach problem solving or "higher order thinking skills" as part of the school curriculum (e.g. Adey et al., 2007; Becker et al., 2012; Greiff et al., 2014; Kuhn, 2009; Martinez, 2000, 2013; R. E. Mayer & Wittrock, 1996). A remarkable communality of these demands is that they are all obviously highlighting and dealing with the importance of domain-general or cross-curricular skills. It is important that students show a general ability to deal with all kinds of problems across domains. Crucially, as was already critically addressed by Adey, et al. (2007, p. 76): "these efforts point in one direction, although they avoid naming the key construction (…): the conception of a general cognitive ability or intelligence." Therefore, and to put it in other words, current trends in education lead to an increased consideration and training of cognitive abilities but using a different label for them.

Based on this reasoning, global, large-scale assessment programs, like the Program for International Student Assessment (PISA), that evaluate the efficiency of educational systems, also increasingly include trans-curricular, content-free domains, such as problem solving, in their assessment framework (cf. Greiff et al., 2014; OECD, 2010). Both related trends, however, implicate the assessment of cognitive abilities. In the first case, formative evaluation of specific courses and related research is needed and the latter asks for summative evaluation of the curriculum's efficiency.

1.2. Traditional tools to assess students‘ cognitive abilities – Intelligence tests

For more than 100 years, education’s challenges to deal with cognitively heterogeneous children has successfully been met by the use of one of psychology’s most important technologies: standardized tests of cognitive abilities, commonly referred to as “intelligence” or “IQ” tests (cf. Hunt, 2011; Kaufman, 2000; Mayer, 2000). Today, there is a remarkable number of such assessment instruments; for a detailed overview, see for example Brickenkamp’s test compendium (Brähler, Holling, Leutner, & Petermann, 2002) or Süß and Beauducel (2011). These tests differ concerning the degree of theoretical reference (almost none vs. well-founded in theory), the type of underlying theoretical framework (e.g., neuropsychology, cognitive psychology, or psychometric approach), as well as under which circumstances they are administered (individually vs. within a group setting). Crucially though, probably as a matter of success, the fundamental design and underlying measurement concept has not changed much since their invention (Hunt, 2011; Sternberg & Kaufman, 1996). For example, nearly all of them encompass tasks that measure numerical, figural-spatial, and verbal skills. Additionally, item formats of the current version of the popular *Stanford-Binet Intelligence Scale*, fifth edition (SB5; Roid, 2003) are still strongly oriented on its predecessor, the Binet-Simon Scale, published in the year 1905 (K. A. Becker, 2003; Nicolas et al., 2013). The same holds true for the equally popular tests in the tradition of the *Wechsler-Intelligence Scales* (Wechsler, 1939), for example the current *Wechsler Intelligence Scale for Children – Fifth Edition* (Wechsler, 2014).

1.2.1. The appearance and psychometric structure of cognitive ability tests

The aforementioned constancy of the measurement concept allows for illustrating the kinds of problems often encountered in cognitive ability tests by presenting three typical item types demonstrating the domains of numerical, figural-spatial and verbal reasoning (Fig. 1).

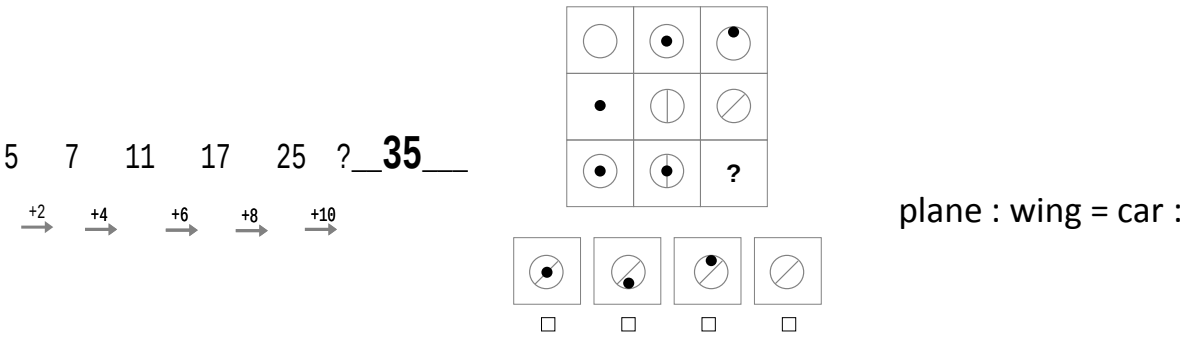


Fig. 1. Typical problems encountered in cognitive ability tests

Cognitive processes involved in solving these item types have been extensively investigated. For example, the first item type, representing numerical reasoning, asks the test taker to complete the given series of numbers according to the rules that can be found in the given numbers. After the test taker detects the relations between the given numbers, the periodicity is discovered. This is followed by completing the pattern description and extrapolating the missing number (cf. Holzman, Pellegrino, & Glaser, 1983; Kotovsky & Simon, 1973). LeFevre and Bisanz (1986) argued that detecting the relations could be split again into three different sub processes: recognition of memorized numerical series, calculation, and checking. Although drawing on different knowledge domains, comparable processes are involved when the other two item types are solved, including the initial identification of certain patterns, an analysis of given relationships and the extrapolation of

these rules to arrive at the solution (concerning an analysis of matrix items, see Carpenter, Just, and Shell, 1990; or Mackintosh and Bennett, 2005; concerning verbal analogies, see Bejar, Chaffin, and Embretson, 1991; or Roccas and Moshinsky, 2003). Hence, the tasks of cognitive ability tests draw on common cognitive processes but also differ in significant aspects.

The question of how to best represent these relations gave rise to various measurement models and was highly disputed, since it evidently also addresses the question of the structure of human cognitive abilities themselves (Schulze, 2005). However, since the impressive factor-analytic studies of Carroll (1993), the view of a hierarchical conceptualization of cognitive abilities has dominated (Deary, 2012; Hunt, 2011; Schulze, 2005). Basically, this means that (statistical) factors of differing generalization or "broadness" have been found. Some explain performance differences in nearly all kinds of tasks within a cognitive ability test (broad or higher-order factors), referring to the notion that if one performs well in one kind of task, he is also likely to perform well in another kind of task. Other, so-called "lower-order" factors were found to explain variance only in a specific set of tasks, thus being "narrower." Given the characteristics of the tasks sharing a common factor, it is possible to infer the kind of cognitive ability that is represented by this factor.

These observations finally led to the formulation of the widely-accepted Cattell-Horn-Carroll (CHC) model of cognitive abilities (McGrew, 2005, 2009), suggesting a three hierarchical structure with a general ability, or *g*-factor, at the top, broad ability factors (such as fluid reasoning or short-term memory) at the second level, and narrow sub-factors (such as quantitative reasoning or memory span) at the first stratum. Recently, Johnson and Bouchard (2005) have suggested an even more pronounced hierarchical structure of cognitive abilities, a *g*-VPR model that proposes a four level structure by again putting a general *g*-factor at the top of the hierarchy and reorganizing some of the broader abilities of the CHC-model. The concept of *g* found additional support when it was shown that the general factors of different cognitive

ability tests are in fact identical (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004). Thus, a hierarchical structure of cognitive abilities comprising *g* is empirically well supported.

1.2.2. What is "intelligence" and is it captured by traditional tests of cognitive ability?

Since the invention of cognitive ability tests, the question of what the tests actually measure has been disputed statistically but also semantically. This largely had to do with the illustrious word "intelligence," which was not only used as an umbrella term to subsume cognitive abilities that were captured by the tests, but also served as a projection screen of what should be measured by them. Hardly surprisingly, "intelligence" had different meanings depending on whether it was used by lay people (mostly referring to a "broad" definition of the ability to be successful in all domains of life) or scientists (more closely referring to the nature of the used tasks to assess it, referring to a "narrow" definition) (cf. Stanovich, 2009). Crucially though, up till now not even in the scientific community has consensus been reached concerning a clear-cut definition of the term, leading to numerous descriptions and theories of intelligence (cf. Hunt, 2011; Hunt & Jaeggi, 2013; Sternberg, Lautrey, & Lubart, 2003).

What comes closest to an accepted definition can be found in a very influential editorial that was intended to represent mainstream thinking on intelligence (Gottfredson, 1997). Fifty-two experts in the field of intelligence research agreed on the definition that "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience. Further, intelligence […] "reflects a broader and deeper capability for comprehending our surroundings – 'catching on,' 'making sense' of things, or 'figuring out' what to do." (Gottfredson, 1997, p.13). Contrary to the results of an earlier survey among experts by Snyderman and Rothman (1987), an additional conclusion of this editorial was that

these abilities can be measured by traditional tests of cognitive ability, or "intelligence tests," quite well.

Thus, at least for the field of psychometric research there seems to be considerable agreement that intelligence can be defined in terms of higher cognitive processes, such as (a) ability to learn and acquire knowledge, (b) problem solving, and of course (c) reasoning. For alternative and also highly popular definitions of intelligence outside the psychometric domain, see for example Gardner (1983, 1999), J. D. Mayer, Salovey, and Caruso (2004), Stanovich (2009), or Sternberg (1999). Note, however, that after initially refusing the psychometric definition of intelligence, later versions of all of these alternative conceptions by and large accepted the psychometric definition but rather suggest augmenting it with other abilities, cognitive, as well as non-cognitive ones.

This dispute impressively points to the fact that a clear distinction is needed between the (manifest, directly observable) measures of cognitive abilities and the (latent, not directly observable) construct(s) that should be measured and that are incorporated into the related definitions (e.g., Borsboom, Mellenbergh, & Van Heerden, 2004; Embretson, 1983, 1998). Thus, in the following, when we refer to the manifest level, we will speak of cognitive ability tests. Instead, when referring to the latent construct of intelligence as defined above, we will use the term General Cognitive Ability (GCA). Note that this term has also been used by Carroll (1993) in his influential psychometric theory on the structural organization of cognitive abilities (see 1.2.1.).

1.2.3. Shortcomings of cognitive ability tests

Although consensus on the meaning of GCA has been reached, as comprising (a) the ability to learn and acquire knowledge, (b) problem solving, and (c) reasoning, scholars were

not always in agreement as to whether it was adequately captured by cognitive ability tests, highlighting a discrepancy between the construct that should be measured and the way this is done. For example, regarding the expert definitions concerning the ability to learn, it is important to note that this facet has been described by experts as "capacity to acquire knowledge" (Snyderman & Rothman, 1987) or "learn quickly and learn from experience" (Gottfredson, 1997, p.13). Therefore it is evident that this facet of GCA goes beyond simple memorizing, especially when it concerns learning "from experience."

The test content of an adequate assessment should therefore enable the test taker to acquire new knowledge and – if possible – experience-driven learning. However, in 1987, 661 experts in the fields of education and psychology participated in a survey by Snyderman & Rothman (1987), and 42% of them indicated that the "capacity to acquire knowledge" is not adequately measured by standard cognitive ability tests. This becomes evident when investigating the typical ways in which cognitive ability tests cover this facet. The first lies in addressing short-term memory (e.g. the SB5, Roid, 2003), typically by subscales which are based on early roots of research on learning, and present single, context-free words or figures, which have to be learnt in a certain amount of time. The second approach to measure this facet lies in testing already acquired knowledge in the verbal or numerical domain. Due to the static nature of paper-pencil tests, however, an interaction between the test taker and the material to be learnt is not possible, thus leaving the test taker's learning strategy as an unknown.

Critique concerning adequate coverage facet problem solving comes from the domain of problem solving research itself (e.g. Dörner, 1986). This research strand showed that using relatively simple structured problems (e.g. Tower of Hanoi or Puzzles like Tangram) with a single, optimal solution failed to capture the basic processes responsible for solving all kinds of (everyday) problems (Funke, 2003; Wenke, Frensch, & Funke, 2005). As a reaction, attention focused on how more complex problems are solved (CPS) by trying to simulate real

world problems using complex computer simulations (for an overview, see 1.3.). CPS is needed "to overcome barriers between a given state and a desired goal state by means of behavioral and/or cognitive, multi-step activities. The given state, goal state, and barriers are complex, change dynamically during problem solving, and are intransparent as well as unknown to the individual problem solver at the outset." (cf. Frensch & Funke, 1995, p.18). Consequently, typical problems or tasks encountered in cognitive ability tests (as can be seen in Fig.1) obviously lack several key characteristics of problems considered to mirror real-world problems. The tasks were therefore found to have a low level of complexity, while also being transparent, static, well-structured, and having only one clearly defined goal (Brehmer & Dörner, 1993; Dörner, 1986; Dörner & Kreuzig, 1983).

Critique concerning insufficient coverage of GCA by cognitive ability tests is not motivated by detailed content-related task analysis alone. On a surface level, it is barely obvious why individuals that successfully complete prototypical reasoning items such as those depicted in Figure 1, are more successful at their future occupations than the individuals who fail to solve them. Thus, despite the undeniable and impressive predictive validity of cognitive ability tests, they were frequently criticized for lacking face validity, which subsequently lead to low acceptance among test users and test takers (Kersting, 1998; Kröner, Plass, & Leutner, 2005; Kyllonen & Lee, 2005).

1.3. A promising alternative to assess cognitive abilities? – Simulated complex problems

Scholars in the domain of complex problem solving (CPS) research have dismissed the typical problems of cognitive ability tests as unrelated to complex real-life tasks. Dating back

to the 1970s, Dietrich Dörner, whose work is regarded as a pioneering effort on CPS, thus wanted to improve the assessment of cognitive abilities by introducing his concept of "operative intelligence" (Dörner, 1986; Dörner & Kreuzig, 1983). By attempting to model and simulate complex everyday problems via computer-based scenarios, he and his associates were trying to overcome the limitations of classical cognitive ability tests (Brehmer & Dörner, 1993; Brehmer, Leplat, & Rasmussen, 1991). Contrary to the impressive constancy of cognitive ability tests, though, the measurement of complex problem-solving ability underwent substantial change during past few decades.

1.3.1. The measurement of complex problem solving

1.3.1.1. Assessing complex problem solving with complex, computer-based scenarios

A typical example of early work on the measurement of complex problem solving is the Lohhausen scenario (Dörner, Kreuzig, Reither, & Stäudel, 1983), a framework which includes more than 2000 interconnected variables. The scenario involves a small village called Lohhausen, which is managed by the test-taker, who must ensure the wellbeing of the town's inhabitants and to guarantee a flourishing economy within 10 (simulated) years. By asking the experimenter questions, the participants could gather information about the system components and then make decisions about them; for example, a test-taker might raise the salary in the city's company or build a new block of flats. A computer program then calculated the impact of the decisions on the system.

The extent to which participants were successful in running the village was measured as a conglomerate of the fictive employees' satisfaction, the employment rate, and/or the production rate of Lohhausen's company. The underlying structure of this scenario allowed for the realization of certain characteristics which were then considered essential for solving

complex everyday problems (e.g. Dörner et al., 1983). In governing Lohhausen, test subjects had to deal with (a) complexity; (b) connectivity between a huge number of variables; (c) dynamic development of certain variables; (d) opacity of the underlying connections; and (e) multiple goals. This approach provided high face validity and was very appealing in the practitioner's eye. It is therefore no surprise that in the context of research or personnel selection a high number of different complex problem-solving scenarios were developed, leading to an "inflation of methods" (Kleinmann & Strauß, 1998).

A drawback to this euphoria concerning the "apparent" resemblance to real life was the fundamental lack of a common theoretical framework, making a comparison between them and/or a conclusion about the construct(s) they aimed to measure impossible (see Frensch & Funke, 1995; Funke, 1993; Funke, 2003; Rollett, 2008). Moreover, manifold methodological concerns were raised (cf. Buchner, 1995; Funke, 1993; Kröner, 2001; Kröner et al., 2005; Süß, 1996, 1999). First, although using semantic labels and more than 2000 variables in modeling the underlying connections, it is highly doubtful that the simulated processes have much in common with reality. If the real-world imitation were successful, prior knowledge would be a clear advantage; otherwise, (correct) prior knowledge would hamper performance in the scenarios since it would lead the participant to rely on wrong assumptions instead of making new experiences while exploring the scenario.

Another methodological problem is the existence of multiple, in most cases ill-defined, goals like "improving life quality," which forces the test taker to prioritize and to set individual standards for goal achievement. Some scenarios consist of a level of complexity that makes it impossible for even test developers to define an optimal solution, thus rendering a standardized assessment of test takers' goal achievement impossible. Finally, since each scenario consists of several discrete time steps, each influenced by the decisions made in the step before, the problem situations encountered are highly dependent on one another. This dependency, though,

violates fundamental assumptions of classical, as well as probabilistic, test theory and can be described as contra-adaptive (Kröner, 2001), meaning that test takers with poor performance are confronted with situations of increasing difficulty during the test session.

1.3.1.2. Complex problem solving scenarios based on formal task analysis

As a reaction to this theoretically and methodologically unsatisfying situation, Funke and associates in particular (e.g., Blech & Funke, 2005; Buchner, 1995; Funke, 1993) promoted a more structured and psychometrically-sound approach based on formal task analyses, which should guarantee comparable complex problem-solving scenarios and results. The focus was set on carefully-constructed scenarios instead of high face validity at any cost. Although largely based on the earlier introduced characteristics to define a complex problem (complexity, connectivity, dynamics, opacity, and multiple goals), these systems used a formalism that offers several advantages. Apart from having clearly-defined optimal solutions (i.e., at any time in the simulation it is possible to determine what a correct intervention would be), these systems facilitated the systematic manipulation of problem characteristics, allowing, for example, the identification of their specific difficulty.

One such formalism is that of linear equation systems (LES). LES relate input (exogenous) variables to output (endogenous) variables through a set of linear equations of the form $Y_{t+1} = A \times Y_t + B \times X_t$, where Y_t and X_t are vectors describing the state of input (X) and output (Y) variables at the current time (step t); the A and B are matrices containing the weights associated with the variables; and Y_{t+1} is the state of the output variables at the next time step (t + 1). This simple formulation can describe (and simulate) systems with direct and indirect effects, the *Eigendynamik* (the effects of variables on themselves), and, by a simple extension, time-delayed effects.

Scenarios based on LES are typically differentiating between an initial exploration phase, where the test takers are instructed to merely gather knowledge about the system, and a second phase of knowledge application in which the participant is instructed to achieve certain goal values of the outcome variables. These two phases allow for deriving indicators of test takers' (a) problem understanding (i.e., system knowledge) and (b) ability to apply their knowledge to achieve certain targets (i.e., control performance). However, scenarios in this tradition also fall short in solving the severe problem of dependency caused by the fact that only one scenario has to be explored and, later on, controlled. For a thorough review of the key findings of related research, see Blech and Funke (2005). Recent examples based upon LES are MultiFlux from Kröner (2001) or ColorSIM from Kluge (2008).

1.3.1.3. Problem-solving scenarios of reduced complexity

The most recent development based on LES is the so-called MicroDYN approach (Greiff, Wüstenberg, & Funke, 2012), which tries to overcome the problem of dependency. By presenting several completely independent scenarios that are reduced in complexity, test takers can work on 8-12 scenarios or items within 1 hour, each item including a phase to explore the system; the possibility to draw the hypothesized mental model; and a control phase in which certain values have to be achieved within a fixed amount of time. Using LES as underlying formalism provides several advantages: (a) a theoretical embedment; (b) the construction of an infinite item pool; (c) item independency and, due to the possibility to use semantic labels and describe everyday activities with this formalism, (d) ecological validity.

Meanwhile, it has been shown that this approach allows for a psychometrically-sound assessment of complex problem-solving behavior within a reasonable amount of time (e.g. Greiff, Wüstenberg, & Funke, 2012). Crucially, the computer-based format permits the

tracking of each of the test taker's interactions with the problem, thus allowing the deduction of process measures of the applied problem-solving strategies. In using causal diagrams, in which the test taker draws the observed relations between the problem's variables, it is also possible to grasp the test taker's mental representation of the problem (Funke, 1992; Leutner, Funke, Klieme, & Wirth, 2005).

1.3.2. Possible benefits of complex problem-solving scenarios for assessing cognitive abilities

In light of the need to assess students' cognitive abilities (see 1.1.), complex problem-solving scenarios seem to answer at least some of the shortcomings of the tools traditionally used for this purpose (see 1.2.3.). First of all, given the theoretical background, such scenarios seem to provide a better operationalization of problem solving than traditional tests of cognitive ability. Instead of static, transparent, and well-structured problems, complex problem-solving scenarios are, by definition, more complex since they are comprised of dynamic changes, a higher number of involved variables whose relations have to be actively explored, and several targets that have to be achieved. Since cognitive ability tests were criticized as insufficiently covering the facet problem solving of GCA, these problem-solving scenarios might be a promising alternative.

The apparently higher complexity and closer resemblance to real-life problems also makes the scenarios more face valid, thus possibly leading to a higher acceptance among educators (Kröner, 2001; Ridgway & McCusker, 2003). This is also reflected by the use of such scenarios to measure problem solving in international large-scale assessments like PISA (Leutner, Fleischer, Wirth, Greiff, & Funke, 2012; OECD, 2010, 2014b). Finally, an additional advantage of complex problem-solving scenarios in the context of education can be seen in the information they provide on the problem-solving process itself. Whereas traditional cognitive

ability tests were mainly administered in a paper-and-pencil format, and thus only provided information on whether a student has succeeded or failed in solving a problem – the final outcome of problem solving – today's computer-based scenarios provide hints where the problem-solving process might have failed. The differentiation between exploring a problem, gathering knowledge about it, and then applying this knowledge to achieve certain targets (see 1.3.1.), would allow for targeted interventions or evaluations of certain training programs with new educational goals (see 1.1.4.).

1.4. The present dissertation

Given the limitations of traditional cognitive ability tests (1.2.3.), the use of complex problem-solving scenarios seems to be a perfect alternative, or at least a promising supplement (1.3.2.). However, compared to a century of profound theoretical considerations and comprehensive empirical research on cognitive ability tests, investigations on complex problem-solving scenarios are relatively recent. More precisely, satisfying the psychometric characteristics of such scenarios have only been achieved within the last few years, and despite the great acclaim within the applied field, research on complex problem- solving has been a rather isolated research realm pursued mostly by German scholars (cf. Frensch & Funke, 1995; Gonzalez, Thomas, & Vanyukov, 2005; Quesada, Kintsch, & Gomez, 2005). Consequently, there is a substantial lack of empirical work supporting the current euphoria within the educational sector to use such scenarios (e.g. Leutner et al., 2012; Ridgway & McCusker, 2003). This is even more true in light of comprehensive studies showing that complex problem-solving performance could to a large degree be explained by performance in old-fashioned cognitive ability tests, thus questioning the scenarios' incremental validity and

suggesting that they might be nothing more than old wine in new skins (Kröner, 2001; Kröner et al., 2005; Süß, 1996).

In short, a sound judgment on the suitability and possible benefits of complex problem-solving scenarios for the assessment of students' cognitive abilities is premature but would be essential. The present dissertation is aimed at filling this gap by (1) developing a state-of-the-art complex problem-solving scenario that is specially adapted for the educational sector (see 2.1.), and (2) investigating its acceptance within the target population of students (see 2.2.). Moreover, we (3) explore its psychometric structure and external validity (see 3.1.), as well as (4) potential benefits for specific educational assessment questions (see 3.2.) to provide a sound scientific base for evaluating the scenario's potential for assessing students' cognitive abilities.

1.4.1. The COGSIM project – answering the specific needs of Luxembourg

The present dissertation is based on the COGSIM project (Assessment of Students' General Cognitive Ability with Computer-Based Complex Problem Simulations) that took place from April 2009 - March 2012 at the University of Luxembourg and which was funded by the Fonds National de la Recherche Luxembourg (FNR/C08/LM/06). Under the supervision of Principal Investigator Prof. Dr. Martin Brunner and Prof. Dr. Romain Martin, a multi-disciplinary team of psychologists and IT developers at the University of Luxembourg, and IT-developers at the Centre de Recherche Public Henri Tudor realized the complex problem-solving scenario called Genetics Lab, conducted several small- and large-scale studies, processed and analyzed the data, and finally scientifically and publicly disseminated related research findings. Prof. Dr. Joachim Funke and Dr. Samuel Greiff of the University of Heidelberg, Germany offered additional professional guidance. As the project's Research Coordinator, Mag. Philipp Sonnleitner was responsible for coordinating the participating team,

conceptualizing and developing the Genetics Lab, organizing and realizing the data collections, conducting related data analysis, and disseminating the studies' findings.

The COGSIM project responded especially to the specific and pressing needs of the Luxembourgish school system. Tracking decisions in Luxembourg (see 1.1.1.), as well as individual educational or vocational counseling (see 1.1.3.), are mostly guided by measures of very specific skill and knowledge domains heavily drawing on verbal abilities. It has been argued that these procedures may lead to an underestimation of immigrant students' cognitive abilities and hence, an underrepresentation of these students in higher school tracks (e.g. (Burton & Martin, 2008; Klapproth et al., 2012).

Since Luxembourg has one of the highest rates of students reporting immigration background worldwide (OECD, 2012, 2014a), a significant number of students will not be educated in an appropriate way, leading not only to detrimental effects on these students' lives and careers but also implicating a severe loss of cognitive potential for the country's economy. This is aggravated by the fact that, during recent years, Luxembourg's economy has faced a profound change, rapidly evolving from a production-centered system, mainly relying on the country's steel industry towards a knowledge-based economy, focusing on the financial sector but also fostering research and development (OECD, 2007). Thus, the demand for a workforce that is capable of learning and quickly adapting to an ever-changing and complex occupational environment has dramatically risen. The present situation has also intensified the pressure on the Luxembourgish school system to assure that every student leaves the school system with adequate cognitive abilities (i.e., problem solving, capacity to acquire knowledge, reasoning skills), or, put differently, high General Cognitive Ability (GCA) (cf. 1.1.4.).

Luxembourg faces a great need for adequate, suitable instruments to assess students' cognitive abilities that focus on domain-general problem-solving skills, not only to make

important educational decisions more fair and precise, but also to evaluate the educational system's impact on such skills. The COGSIM project should exactly answer this need by developing a state-of-the-art and psychometrically-sound complex problem-solving scenario, one that is openly published online to allow educators unlimited and easy access. In so doing, the project was the perfect base to determine the potential of such scenarios compared to traditional cognitive ability tests and thus, to answer the key questions of the present dissertation.

1.4.2. Development of the Genetics Lab

At the start of the COGSIM project in spring 2009, existing problem- solving scenarios were mainly developed and psychometrically evaluated by drawing on (mostly German) university student samples. Notable exceptions that focused on student samples, however, focused on students in higher grade levels (10 or above) in the highest academic track (e.g., Kröner, 2001; Rollett, 2008; Süß, 1996). In addition, claims concerning the high acceptance of problem-solving scenarios within the educational sector lacked empirical proof (e.g., Ridgway & McCusker, 2003). A review of existing problem-solving scenarios at that time (e.g., MultiFlux by Kröner, 2001, or MicroDYN by Greiff and Funke, 2010) quickly made clear that due to their different focus, they were not suited for being administered within samples of lower grade levels (levels 9 and below) with higher cognitive heterogeneity (including non-academic, vocational, and academic tracks). Their degrees of complexity (e.g., including numerical inputs and outputs) and their (old-fashioned and outmoded) design did not correspond to the expectations of today's students, who are extensively exposed to well-designed and supremely modern information and communication technologies (ICT) and are thus often labeled as "digital natives" (Prensky, 2001).

Therefore, the initial goal of the COGSIM project was the development of a timely complex problem-solving scenario suited for lower grade levels and all existing school tracks. Chapter 2 describes the development of the so-called Genetics Lab (GL) by drawing on two related publications. The first, a book chapter to be published by the OECD (Sonnleitner, Keller, Martin, Latour, & Brunner, *in press*), explains the general rationale of the GL's development, including extensive usability testing and an in-depth investigation of today's students' ICT-related characteristics. Besides discussing the implementation of game-like characteristics and the employment of usability studies to accommodate the specific demands of the "net generation," the chapter also reflects on the special challenges concerning the scoring of students' performances on complex problem-solving scenarios. For being practically useful in educational settings, scores provided by the GL should not only be psychometrically sound to allow for reliable and valid score interpretations, but should also make full use of the possibilities computer-based testing has to offer (e.g., the use of digital "traces" of the students), and be easily understandable in order to directly support educational practice.

The second publication, which appeared in *Psychological Test and Assessment Modeling* (Sonnleitner et al., 2012), discusses the advantages of the GL in comparison to previous problem-solving scenarios in great detail. The paper further draws on two samples of 9th graders (of various school tracks) to empirically investigate their acceptance of the GL, and to gather initial results on the psychometric characteristics of the GL's performance scores. We studied the relation between the three measured facets of complex problem- solving behavior (i.e., gathering knowledge, documenting knowledge, applying knowledge; see 1.3.1.) with reasoning ability, to learn more about the GL's construct validity. Furthermore, we examined the performance score's external validity by studying their relation to students' mathematics and science grades.

1.4.3. Psychometric characteristics and potential benefits of the Genetics Lab for assessing students' cognitive abilities

After successfully developing the GL, we thoroughly investigated its psychometric structure, its external and incremental validity, and its potential advantages compared to traditional assessment instruments of students' cognitive abilities. We conducted a large-scale study including a representative sample of $N = 563$ Luxembourgish students enrolled in non-academic and academic tracks in grade levels 9 and 11. Together with the GL, we administered three reasoning scales of an established intelligence test battery (IST-2000R; Amthauer, Brocke, Liepmann, & Beauducel, 2001) and a background questionnaire collecting information about the students' sociodemographic characteristics. In addition, we collected data on students' educational success (i.e., self-reported grades and performance in standardized scholastic achievement tests, such as PISA and the national school-monitoring program ÉpStan).

The major findings of this study are found in Chapter 3, which is composed of two publications tackling questions concerning CPS's psychometric structure and potential added value in the educational context. The first publication appeared in *Intelligence* (Sonnleitner, Keller, Martin, & Brunner, 2013), and focuses on a comprehensive investigation of the structure of complex problem-solving ability (CPS) and its relation to reasoning and educational success. Given the inconclusive previous findings on the structure of CPS, the paper adapts a well-balanced stance concerning its psychometric conceptualization. In analogy to the psychometric structure of general cognitive ability (GCA, see 1.2.2.), we conceptualized CPS as either being hierarchical, including a general CPS-factor at the top or as being a faceted construct, shifting the focus to three (distinct) first-order factors representing students' abilities to explore a problem, depict the gathered knowledge, and consecutively apply the knowledge

to achieve given targets. We continued this even-handed approach when investigating the relation between CPS and reasoning ability.

Contrary to the previous study that investigated the relationship between CPS and reasoning only on the (manifest) level of scale scores (Sonnleitner et al., 2012), we now drew on recent methodological advancements in structural equation modeling (e.g., Eid, Lischetzke, Nussbeck, & Trierweiler, 2003) to directly study associations between the (latent) constructs. A juxtaposition of the psychometric conceptualizations (hierarchical vs. faceted measurement models) of both constructs should substantially add to the existing body of knowledge concerning the relation between CPS and reasoning. Furthermore, by studying CPS's external and incremental validity concerning students' educational success, we aimed at a better understanding of what additional diagnostic information could be gained compared to traditional measures of cognitive abilities.

Previous studies on this topic provided mixed results and were mostly conducted outside the educational domain. Moreover, they addressed different levels of generality (general vs. specific ability factors) in the predictor as well as the outcome variables, not allowing for unambiguous conclusions. As a consequence, we continued our even-handed approach and studied the external as well as incremental validity of CPS when at the same time considering different psychometric conceptualizations of the involved constructs (i.e., GCA, CPS, and educational success). In doing so, we were the first to explicitly consider the hierarchy of cognitive abilities when addressing the relation between CPS, reasoning, and external criteria.

The second paper that drew on data from COGSIM's large-scale study was published in the *Journal of Educational Psychology* (Sonnleitner, Brunner, Keller, & Martin, 2014). In continuing with the aim to respond to the specific needs of Luxembourg's educational sector

(see 1.4.1.), we then investigated whether the GL was fair with respect to students' immigration background and explicitly targeted the question of the potential advantages of problem-solving scenarios for immigrant students. Fairness, in a psychometric sense, does not imply that immigrant students perform equally well as their native peers. Instead, fairness or measurement invariance is a given, as soon as performance differences between groups are only caused by differences in the measured construct (in this case, CPS) and not the measuring instrument itself (in this case, the GL) (cf. Little, 1997; Widaman & Reise, 1997). Previous results on measurement invariance of complex problem-solving scenarios were promising, but focused only on two facets of CPS (knowledge gathering and target achievement) and did not encompass immigration background (Greiff et al., 2013; Wüstenberg, Greiff, Molnár, & Funke, 2014). Reasons why measurement invariance could not automatically be assumed in this context were found, for example, in studies identifying culturally dependent strategies for exploring complex problems (Güss, Tuason, & Gerhard, 2009; Strohschneider & Güss, 1999).

On base of our sample of 299 Luxembourgish ninth graders, including a representative number of immigrant students (n = 127), we investigated measurement invariance with regard to students' immigration background for CPS, again conceptualized as both a hierarchical and a faceted construct. Consecutively, we examined performance differences between the groups, hypothesizing that immigrant students who are largely enrolled in lower academic school tracks might benefit from the novel task demands of the GL. Since conducting experiments in a systematic way – a crucial requirement during the problem-exploration phase of the GL – is hardly, or almost never, taught in Luxembourgish schools (MENFP, SCRIPT, Université du Luxembourg, & EMACS, 2007), performance in the GL might be less affected by students' educational background than are traditional tests of cognitive abilities which were found to be positively influenced by attending a higher school track (M. Becker et al., 2012; Gustafsson, 2008). Thus, the GL might have the potential to identify immigrant "underachievers," who lack

language skills but would in general have the cognitive potential to attend a higher academic track. Taking into account that most immigrant students of our sample were enrolled in the lower academic track, we finally ran two multiple-indicator, multiple-causes models (MIMIC; (Jöreskog & Goldberger, 1975; Muthén, 1989), to control for immigration and educational background at the same time. This procedure should allow for gaining further insights on the possible causes of performance differences between the groups.

In sum, results of both studies that are reported in Chapter 3 should allow a more elaborate judgment of the potential of complex problem-solving scenarios for the assessment of students' cognitive abilities. For a general discussion and evaluation of the findings of Chapters 2 and 3, please see Chapter 4.

References

Adey, P., Csapó, B., Demetriou, A., Hautamäki, J., & Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educational Research Review*, *2*, 75–97.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*, 20–29.

Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A.-M. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, *87*, 85–106.

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

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (Seventh). NJ: Prentice-Hall.

Baeriswyl, F., Wandeler, C., Trautwein, U., & Oswald, K. (2006). Leistungstests, Offenheit von Bildungsgängen und obligatorische Beratung der Eltern: Reduziert das Deutschfreiburger Übergangsmodell die Effekte des sozialen Hintergrunds bei Übergangsentscheidungen? [Achievement tests, openness of school tracks, and obligatory counseling of parents: Is the Deutschfreiburger Übergangsmodell reducing effects of social background on tracking decisions?]. *Zeitschrift Für Erziehungswissenschaft*, *9*, 373–392.

Becker, K. A. (2003). History of the Stanford-Binet Intelligence Scales: Content and Psychometrics. In *(Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin No. 1)*. Itasca, IL: Riverside Publishing.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology*, *104*, 682–699.

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and Psychometric Analysis of Analogical Problem Solving.* New York: Springer-Verlag.

Blech, C., & Funke, J. (2005). Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology. Retrieved from (07.05.2015): http://www.psychologie.uniheidelberg.de/ae/allg_en/forschun/dynamis/dynamis_review_08-2005.pdf

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review; Psychological Review*, *111*, 1061.

Brähler, E., Holling, H., Leutner, D., & Petermann, F. (Eds.). (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests*. Göttingen, Germany: Hogrefe.

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.

Brock, C., & Alexiadou, N. (2013). *Education around the world: a comparative introduction*. London, UK: Bloomsbury Academic.

Brunner, M. (2008). No g in education? *Learning and Individual Differences*, *18*, 152–165.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.

Burton, R., & Martin, R. (2008). L'orientation scolaire au Luxembourg: "Au-delà de l'égalité des chances… le gâchis d'un potentiel humain." In R. Martin, C. Dierendonck, C. Meyers, & M. Noesen (Eds.), *La place de l'école dans la société luxembourgeoise de demain* (pp. 165–186). Bruxelles, Belgium: DeBoeck.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

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

Cattell, R. B. (1971). *Abilities: their structure, growth, and action*. Boston: Houghton Mifflin.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*, 453–482.

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: follow up of the 1932 Scottish Mental Survey. *Intelligence*, *28*, 49–55.

Deißner, D. (2013). *Chancen bilden* [Chances educate]. Wiesbaden, Germany: Springer.

Dick, T. P., & Rallis, S. F. (1991). Factors and Influences on High School Students' Career Choices. *Journal for Research in Mathematics Education*, *22*, 281.

Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit - der Beitrag von Familie und Schule [Educational inequalities - the impact of family and school]. *Zeitschrift Für Erziehungswissenschaft*, *8*, 285–304.

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

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

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, Switzerland: Hans Huber.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CTC(M-1) model. *Psychological Methods, 8*, 38–60.

Embretson, S. E. (1983). Construct validity: Construct representationversus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods, 3*, 300–396.

Engel de Abreu, P. M. J., Gathercole, S. E., & Martin, R. (2011). Disentangling the relationship between working memory and language: The roles of short-term storage and cognitive control. *Learning and Individual Differences, 21*, 569–574.

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

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

Funke, J. (1993). Microworlds based on linear equation systems: A new approach to complex problem solving and experimental results. In G. Strube & K. F. Wender (Eds.), *The Cognitive Psychology of Knowledge* (pp. 313–330). Amsterdam, The Netherlands: Elsevier Science Publishers.

Funke, J. (2003). *Problemlösendes Denken* [Problem solving thinking]. Stuttgart, Germany: Kohlhammer.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. New York: Basic Books.

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

Gottfredson, L. (1997). Mainstream science on intelligence (editorial). *Intelligence, 24*, 13–23.

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

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13*, 74–83.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A New Assessment Perspective. *Applied Psychological Measurement*, *36*, 189–213.

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.

Grigorenko, E. L. (2015). Recent research in the field of giftedness: The field in 30 minutes or less. *Online Educational Research Journal*, *4*, Retrieved from (07.05.2015): www.oerj.org.

Güss, C. D., Tuason, M. T., & Gerhard, C. (2009). Cross-National Comparisons of Complex Problem-Solving Strategies in Two Microworlds. *Cognitive Science*, *34*, 489–520.

Gustafsson, J.-E. (2008). Schooling and intelligence: Effects of track of study on level and profile of cognitive abilities. *Extending Intelligence: Enhancement and New Constructs*, 37–59.

Hallinan, M. T. (1994). Tracking: From Theory to Practice. *Sociology of Education*, *67*, 79.

Hattie, J. (2009). *Visible Learning. A Synthesis of over 800 Meta-Analyses relating to Achievement*. London, UK: Routledge.

Heller, K. A. (1991). *Begabungsdiagnostik in der Schul- und Erziehungsberatung [Assessment of potential in school and educationa counseling]*. Bern, Switzerland: Verlag Hans Huber.

Holzman, T., Pellegrino, J., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*, 603.

Hornung, C., Brunner, M., Reuter, R. A. P., & Martin, R. (2011). Children's working memory: Its structure and relationship to fluid intelligence. *Intelligence*, *39*, 210–221.

Hornung, C., Schiltz, C., Brunner, M., & Martin, R. (2014). Predicting first-grade mathematics achievement: the contributions of domain-general cognitive abilities, nonverbal number sense, and early number competence. *Frontiers in Psychology*, *5*.

Hunt, E. (2011). *Human Intelligence*. New York: Cambridge University Press.

Hunt, E., & Jaeggi, S. (2013). Challenges for Research on Intelligence. *Journal of Intelligence*, *1*, 36–54.

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393–416.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g consistent results from three test batteries. *Intelligence*, *32*, 95–107.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *10*, 631–639.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*, 199–225.

Kaufman, A. S. (2000). Tests of Intelligence. In R. J. Sternberg (Ed.). In *Handbook of Intelligence* (pp. 445–476). Cambridge, UK: Cambridge University Press.

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren [Differential aspects of social acceptance of intelligence tests and problem solving scenarios as instruments for personnel selection]. *Zeitschrift Für Arbeits- und Organisationspsychologie*, *42*, 61–75.

Klapproth, F., Glock, S., Böhmer, M., Krolak-Schwerdt, S., & Martin, R. (2012). School placement decisions in Luxembourg: Do teachers meet the Education Ministry's standards? *The Literacy Information and Computer Education Journal*, *1*, 765–771.

Kleinmann, M., & Strauß, B. (1998). Validity and application of computer-simulated scenarios in personnel assessment. *International Journal of Selection and Assessment*, *6*, 97–106.

Kluge, A. (2008). Performance Assessments With Microworlds and Their Difficulty. *Applied Psychological Measurement*, *32*, 156–180.

Kotovsky, K., & Simon, H. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, *4*, 399–424.

Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation [Intelligence assessment via computer simulations]*. Münster, Germany: Waxmann.

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

Kuhn, D. (2009). Do students need to be taught how to reason? *Educational Research Review*, *4*, 1–6.

Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. W. Engle, *Handbook of Understanding and Measuring Intelligence* (pp. 11–25). Thousand Oaks, CA: Sage.

LeFevre, J.-A., & Bisanz, J. (1986). A cognitive analysis of number-series problems: Sources of individual differences in performance. *Memory & Cognition*, *14*, 287–298.

Leutner, D., Fleischer, J., Wirth, J., Greiff, S., & Funke, J. (2012). Analytische und dynamische Problemlösekompetenz im Lichte internationaler Schulleistungsvergleichsstudien. *Psychologische Rundschau*, *63*, 34–42.

Leutner, D., Funke, J., Klieme, E., & Wirth, J. (2005). Problemlösefähigkeit als fächerübergreifende Kompetenz [Problem solving ability as cross-curricular competency]. In E. Klieme, D. Leutner, & J. Wirth, *Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, the- oretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie* (pp. 11–19). Wiesbaden, Germany: VS.

Little, T. D. (1997). Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues. *Multivariate Behavioral Research*, *32*, 53–76.

Maaz, K., Baeriswyl, F., & Trautwein, U. (2013). Studie: "Herkunft zensiert?" Leistungsdiagnostik und soziale Ungleichheiten in der Schule [Study: "Origin censored?" Performance diagnostics and social inequality in school]. In D. Deißner (Ed.), *Chancen bilden* (pp. 185–305). Wiesbaden, Germany: Springer.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, *33*, 663–674.

Martinez, M. E. (2000). *Education as the cultivation of intelligence*. Mahwah, NJ: Lawrence Erlbaum.

Martinez, M. E. (2013). *Future Bright: A Transforming Vision of Human Intelligence*. Oxford, UK: University Press.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Emotional Intelligence: Theory, Findings, and Implications. *Psychological Inquiry*, *15*, 197–215.

Mayer, R. E. (2000). Intelligence and Education. In R. J. Sternberg (Ed.). In *Handbook of Intelligence* (pp. 519–533). Cambridge, UK: Cambridge University Press.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-Solving Transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: Macmillan.

Mcclain, M.-C., & Pfeiffer, S. (2012). Identification of Gifted Students in the United States Today: A Look at State Definitions, Policies, and Practices. *Journal of Applied School Psychology*, *28*, 59–88.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: theories, tests, and issues* (pp. 136–181). New York: Guilford.

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.

MENFP, SCRIPT, Université du Luxembourg, & EMACS. (2007). *PISA 2006: rapport national Luxembourg*. Luxembourg.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Nicolas, S., Andrieu, B., Croizet, J.-C., Sanitioso, R. B., & Burman, J. T. (2013). Sick? Or slow? On the origins of intelligence as a psychological object. *Intelligence*, *41*, 699–711.

Oakes, J. (2005). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University.

OECD. (2007). *Examens territoriaux de l'OCDE: Luxembourg 2007*. Paris, France: OECD Publishing.

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

OECD. (2012). *Untapped Skills: Realising the Potential of Immigrant Students*. OECD Publishing. Retrieved from (07.05.2015): http://dx.doi.org/10.1787/9789264172470-en

OECD. (2014a). Luxembourg. In *International Migration Outlook 2014* (pp. 276–277). Paris, France: OECD Publishing.

OECD. (2014b). *Pisa 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. Paris, France: OECD Publishing.

Piaget, J. (1969). *Das Erwachen der Intelligenz beim Kinde [The awakening of childrens' intelligence]*. Stuttgart, Germany: Klett.

Preckel, F., & Baudson, T. G. (2013). *Hochbegabung - Erkennen, verstehen, fördern [Giftedness - recognizing, understanding, fostering]*. München, Germany: Beck.

Prensky, M. (2001). Digital Natives, Digital Immigrants Part 1. *On the Horizon*, *9*, 1–6.

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

Ridgway, J., & McCusker, S. (2003). Using Computers to Assess New Educational Goals. *Assessment in Education: Principles, Policy & Practice*, *10*, 309–328.

Roccas, S., & Moshinsky, A. (2003). Factors Affecting the Difficulty of Verbal Analogies. *Applied Measurement in Education*, *16*, 99–113.

Roeder, P. M. (1997). Entwicklung vor, während und nach der Grundschulzeit: Literaturüberblick über den Einfluss der Grundschulzeit auf die Entwicklung in der Sekundarschule [Development prior, during, and after elementary school: literature review on the influence of elementary school on the development in secondary school]. In F. E. Weinert & A. Helmke, *Entwicklungen im Grundschulalter* (pp. p. 405–421). Weinheim, Germany: Beltz/ PVU.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition*. Itasca, IL: Riverside Publishing.

Rollett, W. (2008). *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme [Strategy use, generated information and use of information in exploring and controlling compley, dynamic systems]*. Berlin, Germany: Lit Verlag.

Schaefer, S., Krampe, R. T., Lindenberger, U., & Baltes, P. B. (2008). Age differences between children and young adults in the dynamics of dual-task prioritization: Body (balance) versus mind (memory). *Developmental Psychology*, *44*, 747–757.

Schaie, K. W. (1996). *Intellectual Development in Adulthood*. Cambridge: Cambridge University Press.

Schalke, D., Brunner, M., Geiser, C., Preckel, F., Keller, U., Spengler, M., & Martin, R. (2013). Stability and Change in Intelligence From Age 12 to Age 52: Results From the Luxembourg MAGRIP Study. *Developmental Psychology*, *49*, 1529–1543.

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

Schnabel, K. U., & Schwippert, K. (2000). Schichtenspezifische Einflüsse am Übergang auf die Sekundarstufe II [Social strata specific influences at the transition to secondary school II]. In J. Baumert, W. Bos, & Lehmann, *TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie - mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd 1. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (pp. 261–281). Opladen, Germany: Leske & Budrich.

Schulze, R. (2005). Modeling structures of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of Understanding and Measuring Intelligence* (pp. 241–263). Thousand Oaks, CA: Sage.

Shavinina, L. V. (2009). *International handbook on giftedness.* Amsterdam, The Netherlands: Springer.

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, *42*, 137–144.

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.

Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, *106*, 681–695.

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, *41*, 289–305.

Sonnleitner, P., Keller, U., Martin, R., Latour, T., & Brunner, M. (in press). Assessing Complex Problem Solving in the Classroom: Meeting Challenges and Opportunities. In B. Csapó & J. Funke (Eds.), *The Nature of Problem Solving*. Paris, France: OECD.

Stanovich, K. (2009). *What intelligence tests miss: the psychology of rational thought*. Yale, CT: University Press.

Sternberg, R. J. (1999). A triarchic approach to the understanding and assessment of intelligence in multicultural populations. *Journal of School Psychology*, *37*, 145–159.

Sternberg, R. J., & Kaufman, J. C. (1996). Innovation and Intelligence Testing: The Curious Case of the Dog that Didn't Bark. *European Journal of Psychological Assessment*, *12*, 175–182.

Sternberg, R. J., Lautrey, J., & Lubart, T. I. (2003). Where are we in the field of intelligence, how did we get here, and where are we going? In R. J. Sternberg, J. Lautrey, & T. I. Lubart (Eds.), *Models of intelligence: International perspectives* (pp. 3–26). Washington, DC: American Psychological Association.

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

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

Süß, H. M. (1999). Intelligenz und komplexes Problemlösen. *Psychologische Rundschau*, *50*, 220–228.

Süß, H. M., & Beauducel, A. (2011). Intelligenztests und ihre Bezüge zu Intelligenztheorien (intelligence tests and their relations to theories of intelligence). In L. F. Hornke, M. Amelang, & M. Kersting, *Leistungs-, Intelligenz- und Verhaltensdiagnostik (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd.3*, (pp. 97–234). Göttingen, Germany: Hogrefe.

Tang, M., Fouad, N. A., & Smith, P. L. (1999). Asian Americans' Career Choices: A Path Model to Examine Factors Influencing Their Career Choices. *Journal of Vocational Behavior*, *54*, 142–157.

Tiedemann, J., & Billmann-Mahecha, E. (2007). Zum Einfluss von Migration und Schulklassenzugehörigkeit auf die Übergangsempfehlung für die Sekundarstufe I [On the influence of immigration and school class affiliation on the tracking decision for secondary school I]. *Zeitschrift Für Erziehungswissenschaft, 10*, 108–120.

Wechsler, D. (1939). *Wechsler-Bellevue intelligence scale*. New York: The Psychological Corporation.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children - Fifth Edition*. London, UK: Pearson Assessment.

Wenke, D., Frensch, P. A., & Funke, J. (2005). Complex problem solving and intelligence - empirical relation and causal direction. In R. J. Sternberg & J. Pretz (Eds.), *Cognition and Intelligence: Identifying the mechanics of the mind* (pp. 160–187). Cambridge, UK: Cambridge University Press.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and Assessment in Schools* (Second). New York: Longman.

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, *29*, 18–29.

Chapter II – Development of the Genetics Lab

2.1. Assessing Complex Problem Solving in the Classroom: Meeting Challenges and Opportunities

Sonnleitner, P., Keller, U., Martin, R., Latour, T., & Brunner, M.

(*In press*, to appear in: *The Nature of Problem Solving*, Paris: OECD)

Abstract

At the time when complex problem solving was established as a key aspect of today's educational curricula and a central competence of international assessment frameworks like PISA, it became evident that the educational context places special demands on assessment instruments used for this purpose. In this chapter, we show how these challenges can successfully be addressed by reviewing recent advancements in the field of complex problem solving. We use the example of the Genetics Lab, a newly developed and psychometrically sound microworld, which emphasizes usability and acceptance amongst students, to discuss the challenges and opportunities of assessing complex problem-solving in the classroom.

2.1.1. Introduction

It seems beyond doubt that in a world facing challenges like globalization, global warming, the financial crisis, and depleted resources, the problems our society has to solve will become more complex and difficult during the next years. In their function to prepare younger generations for successfully responding to these enormous challenges, schools have to adapt, too. Therefore, it is not surprising that many contemporary educational curricula and assessment frameworks like PISA (OECD, 2004, 2010) stress the integration and assessment of the ability to solve (domain general) complex and dynamic problems (Leutner, Fleischer, Wirth, Greiff, & Funke, 2012; Wirth & Klieme, 2003). In order to achieve this, many scholars suggest the use of computer-based problem-solving scenarios, so-called "microworlds" that allow for tracking the student's problem-solving process as well as the student's problem representations (Bennett, Jenkins, Persky, & Weiss, 2003; Ridgway & McCusker, 2003) – crucial information for interventions aimed at rising problem-solving capacity in students.

Surprisingly, despite the great enthusiasm about microworlds in the educational field, most previous studies have drawn on adult samples, typically of high cognitive capacity (e.g., university students of various branches). Only a few studies have directly applied such microworlds and investigated their psychometric properties in populations of school students so far. These exceptions, however, mainly focused on students of the higher academic track, and usually at grades 10 or above (e.g., Kröner, Plass, & Leutner, 2005; Rigas, Carling, & Brehmer, 2002; Rollett, 2008; Süß, 1996). Thus, due to the highly selective samples of these studies, it is questionable to what extent microworlds can unconditionally be applied to the whole student population without modifications of their construction rationale or scoring procedures.

This chapter identifies and discusses challenges that arise when microworlds are administered "in the classroom": the special characteristics of today's students, also described

as “digital natives,” and the need for timely, behaviour-based scoring procedures that are at the same time easy to understand by educators and teachers. By taking the Genetics Lab, a microworld especially targeted at students at age 15 and above of all academic tracks as an example (Sonnleitner et al., 2012a), opportunities to react on these challenges are presented and evaluated on the basis of three independent studies using the Genetics Lab.

2.1.2. The Genetics Lab: a microworld especially developed for the educational field

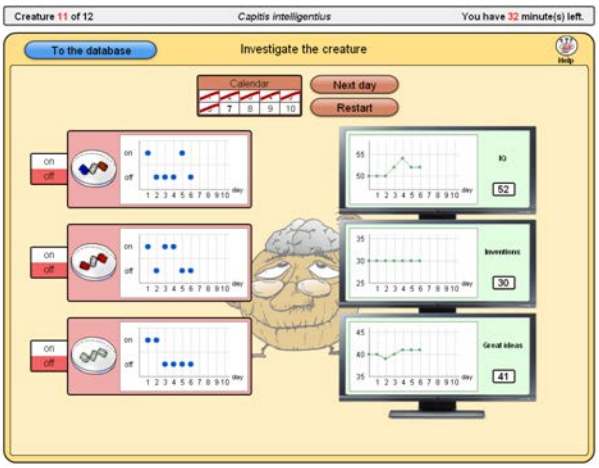
To learn more about the application of microworlds in the educational setting and to further investigate to what extent and in what way microworlds have to be adapted for this context, the Genetics Lab (GL) was developed at the University of Luxembourg (see Sonnleitner et al., 2012a, 2012b). The goal was to set up a face valid, psychometrically sound microworld to assess complex problem solving (CPS) that can immediately be applied in the school context. To this end, the development drew on the rich body of empirical knowledge that was derived from previous studies on microworlds (for an overview, see for example Blech & Funke, 2005; Funke & Frensch, 2007). To enable educators to make full use of the GL, it can be administered within 50 minutes (i.e., the length of a typical school lesson), and in three different languages (English, French, and German). Moreover, it was published under open-source license and can be freely downloaded and applied.[1]

In the GL (shown in Fig. 2), students explore how genes of fictitious creatures influence their characteristics. To this end, students can actively manipulate the creatures’ genes by switching them “on” or “off” and then studying the effects of these manipulations on certain characteristics of the creatures (Fig. 2, a). Genes (i.e. input variables) are linked to the characteristics (i.e. output variables) by linear equations. It is the task of the student to find out

[1] See http://www.assessment.lu/GeneticsLab for additional information and the GL download.

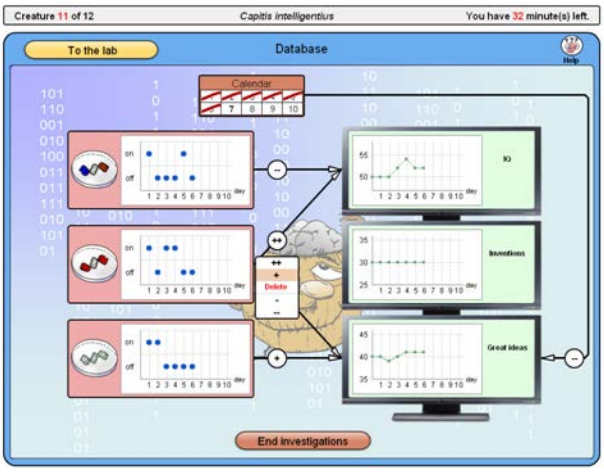
about these (non-transparent) relations and to document the gathered knowledge (Fig. 2, b). Finally, the students have to apply the gathered knowledge to achieve certain target values on the creatures' characteristics (Fig. 2, c). These task characteristics allow for deriving performance scores about (a) the students' exploration and information-gathering behavior; (b) the students' gathered knowledge in the form of a causal diagram showing the discovered relations between genes and characteristics; and (c) the students' ability to apply the knowledge in order to achieve certain target values on the creatures' characteristics.

Each creature is designed in such a way as to realize key features of a complex problem (see e.g., Funke, 2001; Funke, 2003): (a) *complexity*, by including a high number of variables (several genes and characteristics); (b) *connectivity*, by linking the variables via linear equations; (c) *dynamics*, by implementing an automatic change of certain characteristics that is independent from the students' actions; (d) *intransparency*, by hiding the connections between the variables; and (e) *multiple goals*, by asking the student to achieve different target values on several of the creature's characteristics. For further details about the GL's scores and construction rationale please see (Sonnleitner et al., 2012a). The GL has been applied in three independent studies so far with more than 600 participating students (see Table 1 for an overview). To foster commitment and motivation, detailed written feedback on the performance was offered. Further details concerning Study 1 and 2 are given in (Sonnleitner, et al., 2012a), concerning Study 3 in (Sonnleitner et al., 2012b). The gathered data along with the experiences made within these studies inform the following discussion of challenges and opportunities of microworlds within the educational field.



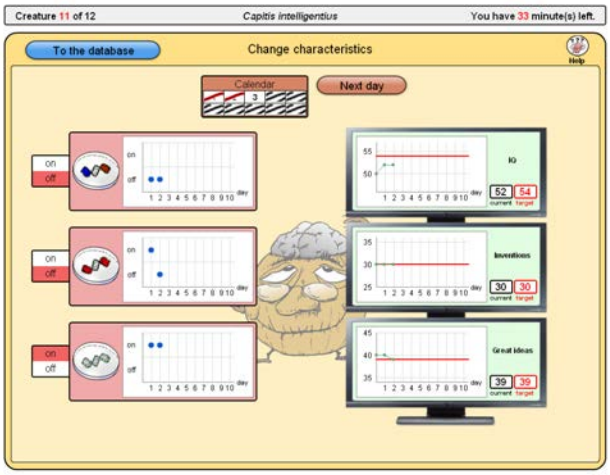
a.: Gathering Knowledge

Students begin with gathering knowledge on how certain characteristics of a creature are affected by its genes. To this end, they switch genes and thus their effects "on" or "off" and then study the consequences of their manipulations in the related diagrams (genes are depicted in left, characteristics are depicted in right diagrams).



b.: Documenting Knowledge

Knowledge that has been gathered about the genes' effects can be documented in a related database that shows the same genes and characteristics as the lab. Students can depict their mental model of the relations by drawing a causal diagram that also indicates the strength and direction of the discovered effects.



c.: Applying Knowledge

In the final phase, students have to achieve certain target values on the creature's characteristics by applying their gathered knowledge. Importantly, they have to accomplish this goal with a limited amount of manipulations. Thus, students have to anticipate potential dynamics of the problem and to plan their actions in advance.

Fig. 2: Screenshots of the different phases students go through within the Genetics Lab (taken from Sonnleitner et al., 2012a)

Table 1: Sample characteristics of the presented studies

		Study 1	Study 2	Study 3
n		43	61	563
Mean age (SD)		15.8 (.87)	15.5 (.61)	16.4 (1.16)
Male		24	26	279
Female		19	35	284
School track	**intermediate**	43	35	234
	academic	-	26	329
School grade	**9th**	43	61	300
	11th	-		263

2.1.3. Challenge 1 – Digital natives

A crucial aspect of an assessment instrument is its suitability for the characteristics and background of the target population (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999, p.131). Previous studies using microworlds mostly drew on adult samples, typically university students. Thus, the question arises whether and in what way today's students differ from these homogenous samples.

Today's students were born in the late 1990s and are described as members of the "net generation" (Tapscott, 1998) or are even called "digital natives" (Prensky, 2001), mainly because they have grown up in a world in which information and communication technology (ICT) is permanently available. Compared to former generations, they deal with digital media like video games, simulations, the internet, instant messaging, virtual learning environments, and social networks on a daily basis. Hence, some authors have claimed that this interaction with digital media right from birth has caused today's students to be cognitively different from prior generations.

According to Prensky (2001), digital natives think of and process information in a manner that is fundamentally different from their forebears. They are used to processing information very fast and to applying multitasking to achieve their goals. They strongly rely on graphics and symbols to navigate and due to their exposure to video games, they are used to getting instant gratification and frequent rewards. In a review on information-seeking habits of this generation, Weiler (2005) describes these students as primarily visual learners that prefer to actively engage in hands-on activities instead of passive learning. Veen & Vrakking (2006) highlight the iconic skills of this generation (use of symbols, icons, and colour-coding to navigate within digital environments) that have been developed in order to deal with a massive and permanent information overload. Moreover, technology is perceived in a new way, as being merely a tool for various purposes and one that has to work flawlessly.

Recent reviews, however, have shown that differences between today's students and former generations may be overstated (e.g. Bennett & Maton, 2010). Indeed, several studies showed that this generation is a very heterogeneous sample with varying degrees of digital competence (Li & Ranieri, 2010) and technology use (Margaryan, Littlejohn, & Vojt, 2011). Nevertheless, the same studies report that almost every member of this generation uses a mobile phone, a personal computer or laptop, and has access to the internet. Thus, while claims concerning the cognitive uniqueness and homogeneity of this generation may be exaggerated, virtually nobody questions the heavy exposure and use of digital media and devices of today's students.

This, in turn, has several crucial implications concerning the expectations of today's students with regard to a computer-based test. First, due to their massive exposure to high quality (commercial) computer programs, they expect a perfect and flawlessly functioning technology. Second, especially on the basis of the experiences made with video games and newer mobile devices, a completely intuitive graphical user interface (GUI) is expected.

Students do not want to invest time and effort to figure out how to interact with a program. Third, this GUI should also be appealing and resemble modern standards of design to ensure that the test is perceived as being attractive and of high quality. Fourth, students want to learn how to deal with the task by actively exploring and interacting with it. In contrast, extensive written instructions are very likely to be skipped over by them. Finally, motivation to interact with a test will be high when they get instant gratification or at least instant feedback on their performance. If these criteria are not met by a computer-based test, acceptance of the instrument might be at stake.

2.1.3.1. Responding to the digital natives' needs: game-like characteristics and usability-studies

A first step to responding to these special characteristics of today's students concerned the theoretical conceptualization of the GL. To start with, we ensured a clear and intuitive GUI by following recommendations of user interface design (e.g. Fulcher, 2003). As can be seen in Fig. 2, the structure of the GUI clearly resembles the navigation within the GL. The layer at the top shows the progress within the test itself by indicating the number of creatures (i.e., items) that are left and the remaining time for investigating them. The next layer corresponds to navigating within each item; it provides the buttons to switch between the lab and the database and contains the help function. Finally, all elements to directly manipulate the creature or depict the gathered knowledge about it are arranged within the inner layer of the GUI. In addition, elements belonging together (e.g., the calendar and the buttons to progress in time) share the same colour. To make the design of the GL even more appealing and to increase the motivation to work on the test, we implemented several game-like characteristics (see McPherson & Burns, 2007; Washburn, 2003; Wood, Griffiths, Chappell, & Davies, 2004): A "cover story" was created, putting the student into the role of a young scientist that starts working in a fictive genetics lab. An older scientist charges the student with the mission of investigating several newly discovered creatures and explains the functioning of the lab.

Throughout the test, this “virtual mentor” remains present in the form of an integrated help function. In addition, the fictitious creatures are depicted in a funny cartoon-like style and carry humorous characteristics (Fig. 2 and 3). Hence, after exploring and manipulating the creature, the student gets performance-contingent feedback in the form of two simple scales scoring the depicted causal diagram of phase 1 and the student’s control performance of phase 2 (Fig. 3).

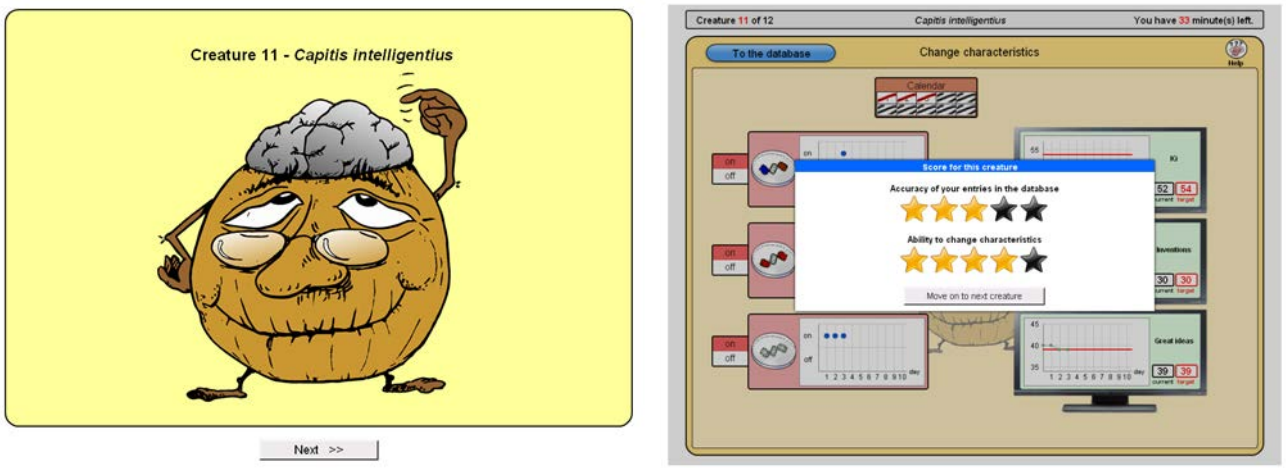


Fig. 3: Start screen of creature (i.e., item) 11 (left side) and performance contingent feedback at the end of an item (right side).

As a reaction to the digital natives’ style of learning, we also laid special emphasis on the instructions given at the beginning of the GL. Whereas former microworlds mostly included extensive written instructions or training periods with varying levels of standardization (Rollett, 2008), the GL’s instructions are highly standardized, interactive, and refer to standards for modern multimedia learning (Mayer, 2003). After a short explanatory text, each task of the GL (exploring the creature, drawing a causal model, and achieving target values) is visualized by an animation and has to be practiced in a related exercise. For drawing the causal model and achieving the target values, detailed feedback about the performance is provided.

The second step to ensure acceptance of the GL among digital natives was to guarantee a flawless functioning test of high usability. To this end, we adapted and substantially extended traditional test development procedures by including several small-scale usability studies (Fig 4). This approach not only aimed at evaluating the design of the GL's GUI in terms of acceptance but also at reducing construct-irrelevant variance in the GL's performance scores (Fulcher, 2003). Participants of the first and second usability studies were experts in the field of testing and usability as well as laypersons. The sample of the third usability study consisted of university students and students of the target population. All participants were asked to think aloud while working on the GL. Together with these comments, the behaviour of the participants was documented by trained observers, and followed by an interview asking participants for perceived problems and possible solutions. On the basis of these data, comprehensibility and functionality problems were identified and discussed in a focus group preparing suggestions for the modification of the GL. The identified problems ranged from minor problems like a suboptimal position of a button to construct-related problems. For example it turned out that using the causal diagram as knowledge representation is highly demanding and unfamiliar to fifteen year-old students.

Whereas results of the first two usability studies caused major revisions of the GUI and especially a modified wording and sequence of the instructions, results of the third usability study merely led to minor changes. Importantly, this approach not only warranted high usability of the GL but also led to substantial insights concerning the measured construct and how to derive valid scoring algorithms of students' problem-solving behaviour.

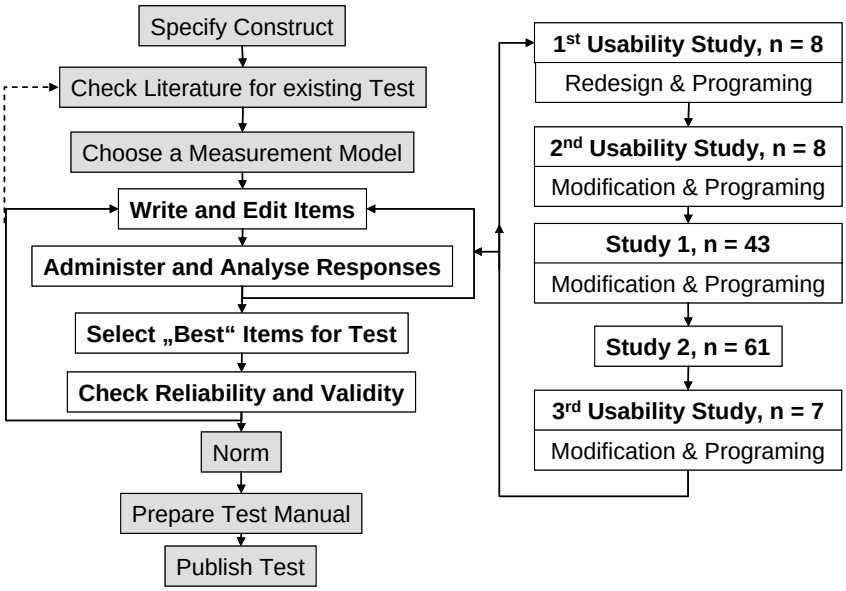


Fig. 4: Adapted test development process of the GL based on the traditional approach presented by Shum, O'Gorman, & Myors (2006)

2.1.3.2. Acceptance and usability of the Genetics Lab among digital natives

A first analysis of our samples' characteristics showed that the claims made by the literature about today's students were well supported by our studies. Figure 5 presents several ICT-related characteristics of the participating students. As can be seen, the vast majority of students have already been using personal computers for more than 3 years (92%) and nearly every day (up to 80%). Moreover, these students report a high ICT-competence on a 10-item questionnaire including several ICT-related activities like burning CDs, downloading pictures and programs from the internet, and creating a webpage or a multimedia presentation. Total scores of this scale were (linearly) transformed into percentage of a maximum possible score (POMP, see Cohen, Cohen, Aiken, & West, 1999). Thus, the percentages depicted in the black bars of Figure 5 (75% for Study 1 and 78% for Study 2) describe the mean achieved percentages of a maximum achievable ICT-competence score.

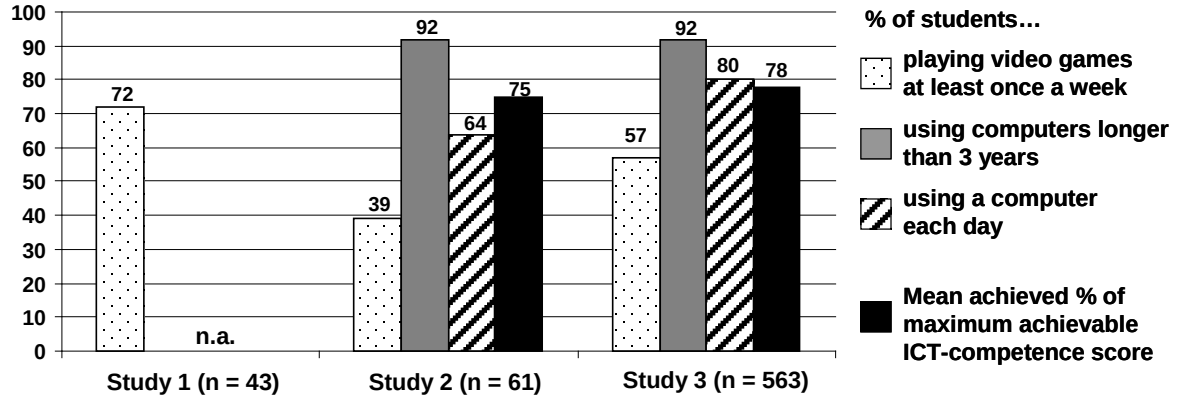


Fig. 5: ICT characteristics of students (n.a. is not applicable)

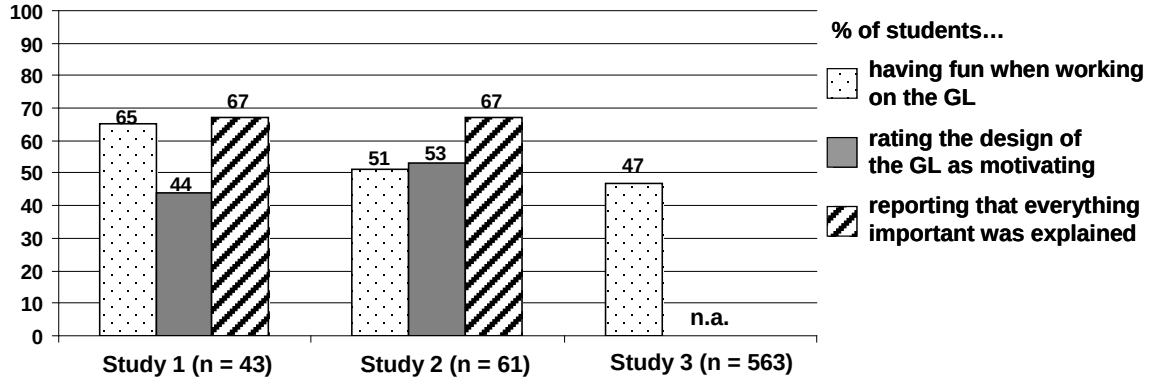


Fig. 6: Acceptance of the Genetics Lab among students (n.a. is not applicable)

Although results concerning the frequency of video game playing per week are somewhat mixed, looking at the most representative Study 3 indicates that about 57% of the participating students play video games at least once a week. Thus, despite a minority that doesn't use computers on a regular basis and rates itself as less ICT-competent, the vast majority of students can be described as ICT-literate with an extensive experience in dealing with digital environments and computer programs.

To investigate whether our attempt to develop a microworld suited for today's students was successful, we evaluated acceptance of the GL within the conceptual framework of well-established technology acceptance models (e.g., Terzis & Economides, 2011). According to this framework, acceptance of a computer-based assessment instrument may be substantially influenced by its *Perceived Ease of Use* and *Attractivity*. Moreover, the instrument's *Comprehensibility* and *Functionality* are considered as crucial factors in determining its usability, and hence its acceptance, among the target population. Consequently, students participating in Study 1 had to rate various elements of the GL in terms of these four dimensions. Results of this questionnaire are presented as POMP-scores in Table 2. Given the lack of comparable studies, we considered values above 50% - indicating that positive student evaluations outweigh negative evaluations - as positive outcomes (for more details about the questionnaire, please refer to (Sonnleitner, et al., 2012). Overall, results show that the GL is well accepted among today's students. The GL was rated as being easy to use (M = 54, SD = 23) with an attractive appearance (M = 64, SD = 22). Students also found the GL to be comprehensible (M = 61, SD = 17) and well-functioning (M = 60, SD = 22). Moreover, in all three studies, large portions of the students reported having fun while working on the GL (see Fig. 6). Apart from Study 1 in which the GL was solitary administered, students had to work on the GL at the end of a 2 hours test session. This may explain the somewhat smaller portion of students in studies 2 and 3 that indicated to have fun while working on the GL. Moreover, results show that the game-like design of the GL was appreciated and even described as motivating by large portions of the students. Crucially, the vast majority of students felt that everything important was explained during instructions. We take this as clear indication of the instruction's efficiency to successfully illustrate the handling of the GL in an interactive multimedia-based way.

To sum up, results suggest that our attempt to develop a microworld of high usability that enjoys high acceptance among today's students was successful. For the first time, we could also go beyond anecdotic evidence that interacting with such scenarios makes fun (e.g. Ridgway & McCusker, 2003). We largely attribute this positive outcome to the actions we have taken to consider special characteristics of today's students, namely, the integration of game-like characteristics, the development of a standardized and interactive multimedia instruction and extensive usability testing.

2.1.4. Challenge 2 – Scoring

The major reasons for using microworlds in the educational context can be seen in (a) assessing and evaluating students' initial CPS-skills and to study their relation to other constructs like general school achievement (Leutner et al., 2012; Wirth & Klieme, 2003), and (b) in directly implementing them into educational practice to use them for interventions aiming at improving these skills (Bennett et al., 2003; Ridgway & McCusker, 2003). This, in turn, poses special challenges concerning the scoring of students' performance on these problem-solving scenarios.

First, and above all, the yielded scores have to be psychometrically sound to allow for reliable and valid score interpretations. Second, in order to be useful for behaviour-based interventions, scores must make full use of the possibilities computer-based testing has to offer. Compared to traditional, mostly paper-pencil-based multiple-choice tests, microworlds allow for capturing the digital "traces" left by the student when interacting with these scenarios (i.e., each action of the student is stored in a related log-file). Although such traces are highly valuable information about the students' problem-solving behaviour, the scoring of such complex behavioural data is challenging (Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007; Winne, 2010). Third, if microworlds are directly used in educational practice to foster

CPS-skills, it is essential that the interpretation of the performance scores yielded be easy and comprehensive. Educators working with these scores should be able to easily understand and use them for drawing sound conclusions about the students' behaviour when confronted with complex problems.

In order to guarantee highly reliable performance scores for the GL, we based its development on the so-called MicroDYN approach (Greiff, Wüstenberg, & Funke, 2012). In contrast to former microworlds that consisted of one very extensive problem-solving scenario, each student has to work on several, independent scenarios (i.e., the creatures) of varying complexity and content. Thus, when students' performance is aggregated across creatures, the resulting scores of the students' CPS-skills are more reliable than those derived from one single scenario. In the following, the GL's performance scores will be discussed concerning their reliability, internal validity, and how the students' traces were used to fully mine the potential of behaviour-based data and to make score interpretation comprehensible.

2.1.4.1. Students' exploration behaviour

In order to gather knowledge about the effects of a creature's genes on its characteristics, students have to explore it by actively manipulating the genes (see Figure 2, a). These manipulations, however, are most informative if students switch one gene to "on" and all other genes to "off". Only then can occurring changes on the creature's characteristics be unanimously interpreted as effects of the gene that is switched "on" (Vollmeyer, Burns, & Holyoak, 1996). Moreover, in order to detect dynamics within a creature (i.e., some characteristics change without being affected by genes), all genes have to be switched "off." Thus, when looking at the students' traces, such informative steps can be distinguished from non-informative ones. This behavioural information can then be used to relate the number of informative steps to the total number of steps taken in the exploration phase across all creatures,

resulting in a behaviour-based *Systematic Exploration* score (Kröner et al., 2005). A high proportion of informative steps thus indicate a student's very efficient exploration strategy.

As can be seen in Table 2, the internal consistency of this performance score is very high, ranging from .88 to .94. The score's validity is further supported by its substantial correlation to the students' gathered *System Knowledge*. The more efficiently a student explores the creatures, the higher his knowledge about them. Interpretation of the score is rather easy, with a theoretical range of 0 to 100, with 100 indicating that the student has only applied informative steps.

To ease and support the use of the GL in educational practice, additional information about the student's exploration behaviour is provided within the published GL-package[1]. Besides the efficiency of the applied exploration strategy, educators can investigate whether the effects of all genes have been investigated; that is, whether all possible informative steps have been realized by the student. This may be valuable information for interventions aimed at students that explore highly efficiently but not conscientiously.

2.1.4.2. Students' gathered knowledge

The students' gathered knowledge about a creature was scored on basis of the causal diagrams which were depicted by the students in the GL's database (see Figure 2, b). These causal models can be interpreted as the theoretical model a student has developed about a creature and are thus valid indicators of his mental problem representations (Funke, 1992). Although the method of knowledge assessment by causal diagrams was often successfully used in samples of university students (Blech & Funke, 2005; Funke, 2001), critique was raised that due to its high cognitive demand, it might be problematic in samples of lower cognitive capacity. This notion, in fact, was supported by the results of our usability studies, which showed that many students of our target population reported problems when using causal

diagrams for knowledge representation. Although we tackled this problem through a substantial modification of the related instructions, it still had to be confirmed that the analysis of causal diagrams yields reliable and valid scores of students' problem representation in our target sample.

For scoring the resulting causal diagrams, we applied a well-established algorithm that differentiates between relational knowledge (i.e., if a relation between a gene and a characteristic exists or not) and knowledge about the strength of these relations (Funke, 1992; Müller, 1993). The student's model is compared to the true underlying relationships and the more similar they are, the higher the knowledge scores that are yielded. Both kinds of knowledge are scored separately and then weighted in order to compose a total *System Knowledge* score. In line with previous studies, relational knowledge was emphasized by multiplying it with a weight of .75, compared to a weight of .25 for knowledge about the strength of an effect (Funke, 1992).

Table 2 shows that the resulting score about the students' gathered knowledge is highly reliable, with a Cronbach's alpha ranging from .77 to .90. Descriptives of this score are given as achieved percentage of a maximum score (POMP, see above). Moreover, the pattern of intercorrelation between *System Knowledge* and *Systematic Exploration*, as well as *Control Performance* supports internal validity of the score: A more efficient exploration strategy leads to higher *System Knowledge* and the higher the gathered knowledge, the better the ability to achieve the target values. Thus, *System Knowledge* can be seen as a reliable and valid measure of students' mental problem representations. In addition to the total *System Knowledge* score, the published GL package also includes both specific knowledge scores: the students' gathered relational knowledge and knowledge about the strengths of effects. To ease score interpretation for educators, the GL's manual contains theoretical minima as well as maxima for each score.

2.1.4.3. Students' control performance

For scoring student's ability to apply the gathered knowledge and achieve certain target values on the creatures' characteristics (see Fig. 2, c), we again drew on behavioural data to compute a process-oriented *Control Performance* score. In order to achieve the given target values within three steps, students have to (a) rely on their knowledge to plan their actions and to forecast possible consequences, and (b) react to unexpected consequences and try to correct them. Both skills are key characteristics of CPS (Funke, 2003). Most previous attempts to score control performance emphasized the (aggregated) deviation between the achieved values and the target values (Blech & Funke, 2005). This approach, however, was criticized for making the scoring of a step dependent on the previous one if the scenario does not allow a participant to reach the target value within one step. A suboptimal step would automatically lead to a deviation from the target value that could not be compensated by the following step. To put it differently, a high skill in correcting problems could not compensate for bad planning behaviour. Consequently, we developed a scoring algorithm that is exclusively based on the students' inputs and that scores every step independently. Only if a step is optimal in the sense that the difference to the target values is maximally decreased, the step is seen as indicating good control performance. Thus, for each creature a maximum score of three is possible.

Internal consistency of the resulting *Control Performance* score was generally acceptable (see Table 2). Though, in Study 2, Cronbach's alpha was rather low, indicating that the mixture of interacting with the creature through reacting and correcting current states may make the scoring of students' control performance not that simple. Nevertheless, results of the most representative Study 3, together with the meaningful pattern of intercorrelations throughout all studies – high *System Knowledge* leads to better *Control Performance* – suggest the score's validity. In addition to the number of optimal steps taken by the student, educators also find the concrete sequence of steps within the GL's package. Interventions therefore could either

target students that lack planning skills given a suboptimal first step or students that show poor control behaviour.

2.1.5. Summary and Outlook

It has been shown that the assessment of complex problem-solving "in the classroom" poses special demands on the assessment instruments used for this purpose. The development of the Genetics Lab successfully responded to most of these challenges by drawing on game-like characteristics, a user interface of high usability, and psychometric sound, behaviour-based scores that are at the same time comprehensive for educators. However, several questions remain to be answered. First, although the vast majority of students in our studies showed characteristics of being "digital native" and thus were likely to be highly competent in using computers and digital media, a minority of students remains that report a low ICT self-competency and only occasional use of modern media. To what extent these students are disadvantaged by the computer-based assessment of their problem solving skills has to be further investigated. Still, given that most future problems of high complexity will be solved in a digital environment this may not be a shortcoming of the assessment instrument but instead contribute to its external validity. Second, although the implementation of game-like characteristics leads to a high acceptance of the GL among today's students, these features could interfere with the measured construct in making the presented problems especially interesting and attractive for some, but not for others. Hence, studies investigating the GL's concurrent validity with other measures of problem solving are therefore needed. Finally, although the scores provided by the GL proved to be internally valid and reliable, their usefulness has yet to be demonstrated in studies that use them for evaluations or interventions. The use of behavioural data is still in its infancy and could substantially benefit from such experiences. In developing the Genetics Lab in three different languages and making it freely accessible online[1], a first step is made to answer these upcoming challenges.

Table 2: Means, standard deviations, reliability, and intercorrelations of the Genetics Lab's

	No. of items	α	M	SD	Min	Max	p25	MD	p75	SE	SK	CP
Study 1 (*n* = 43)												
Complex problem solving												
Systematic Exploration	16	.94	21	12	1	61	13	21	27	1		
System Knowledge	16	.89	54	12	38	96	46	51	57	**.54**	1	
Control Performance	16	.79	32	7	16	47	26	31	35	.27	**.43**	1
Acceptance & usability												
Perceived Ease of Use	4	.71	54	23	0	100	44	56	69	.31	**.44**	**.39**
Attractivity	9	.91	64	22	0	100	56	67	78	.22	**.34**	**.54**
Comprehensibility	10	.81	61	17	20	100	50	63	73	.28	**.49**	**.32**
Functionality	7	.82	60	22	0	100	46	64	71	.17	**.35**	**.50**
Study 2 (*n* = 61)												
Complex problem solving												
Systematic Exploration	12	.88	26	11	7	66	19	25	32	1		
System Knowledge	12	.77	53	12	35	100	45	51	59	**.35**	1	
Control Performance	12	.54	21	4	11	31	18	21	24	**.32**	**.47**	1
Study 3 (n = 563)												
Complex problem solving												
Systematic Exploration	12	.91	28	15	01	71	17	26	39	1		
System Knowledge	12	.90	69	17	37	100	55	67	81	**.55**	1	
Control Performance	12	.79	20	6.8	6	36	14	18	24	**.51**	**.77**	1

performance scores; α = Cronbach's alpha; p25 = first quartile (Q1); p75 = third quartile (Q3)
Note. All coefficients printed in bold are significant at $p < .05$ (2-sided significance).

Acknowledgements

This work was supported by funding from the National Research Fund, Luxembourg (FNR/C08/LM/06). The authors would like to thank all the students and teachers participating in our studies.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Bennett, S., & Maton, K. (2010). Beyond the„digital natives' debate: Towards a more nuanced understanding of students" technology experiences. *Journal of Computer Assisted Learning*, Vol. *26, No.* 5, pp. 321–331.

Bennett, Sue, Maton, K., & Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, Vol. *39, No.* 5, pp. 775–786.

Blech, C., & Funke, J. (2005). Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology. *URL (31.07. 2006): http://www. die-bonn. de/esprid/dokumente/doc-2005/blech05_01. pdf*.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, Vol. *34, No.* 3, pp. 315–346.

Elliot Bennett, R., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing Complex Problem Solving Performances. *Assessment in Education: Principles, Policy & Practice*, Vol. *10, No.* 3, pp. 347–359.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, Vol. *20, No.* 4, pp. 384–408.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & reasoning*, Vol. *7, No.* 1, pp. 69–89.

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

Funke, Joachim. (2003). *Problemlösendes Denken [Problem solving thinking]*. Stuttgart, Germany: Kohlhammer.

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

Greiff, S., Wustenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A New Assessment Perspective. *Applied Psychological Measurement*, Vol. *36, No.* 3, pp. 189–213.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, Vol. *2, No.* 2-3, pp. 107–124.

Jones, C., Ramanau, R., Cross, S., & Healing, G. (2010). Net generation or Digital Natives: Is there a distinct new generation entering university? *Computers & Education*, Vol. *54, No.* 3, pp. 722–732.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, Vol. *33, No.* 4, pp. 347–368.

Leutner, D., Fleischer, J., Wirth, J., Greiff, S., & Funke, J. (2012). Analytische und dynamische Problemlösekompetenz im Lichte internationaler Schulleistungsvergleichsstudien [Analytic and dynamic problem solving competence in the light of international comparative student assessment studies]. *Psychologische Rundschau*, Vol. *63, No.* 1, pp. 34–42.

Li, Y., & Ranieri, M. (2010). Are 'digital natives' really digitally competent?-A study on Chinese teenagers. *British Journal of Educational Technology*, Vol. *41, No.* 6, pp. 1029–1042.

Margaryan, A., Littlejohn, A., & Vojt, G. (2011). Are digital natives a myth or reality? University students' use of digital technologies. *Computers & Education*, Vol. *56, No.* 2, pp. 429–440.

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and instruction*, Vol. *13, No.* 2, pp. 125–139.

McPherson, J., & Burns, N. R. (2007). Gs Invaders: Assessing a computer game-like test of processing speed. *Behavior research methods*, Vol. *39, No.* 4, pp. 876–883.

Müller, H. (1993). *Komplexes Problemlösen: Reliabilität und Wissen. [Complex problem solving: reliability and knowledge]*. Bonn, Germany: Holos.

Prensky, M. (2001a). Digital Natives, Digital Immigrants Part 1. *On the Horizon*, Vol. *9, No.* 5, pp. 1–6.

Prensky, M. (2001b). Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently? *On the Horizon*, Vol. *9, No.* 6, pp. 1–6.

Ridgway, J., & McCusker, S. (2003). Using Computers to Assess New Educational Goals. *Assessment in Education: Principles, Policy & Practice*, Vol. *10, No.* 3, pp. 309–328.

Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, Vol. *30, No.* 5, pp. 463–480.

Rollett, W. (2008). *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme [Strategy use, generated information and use of information in exploring and controlling compley, dynamic systems]*. Berlin, Germany: Lit Verlag.

Shum, D., O'Gorman, J., & Myors, B. (2006). *Psychological Testing and Assessment*. South Melbourne, Australia: Oxford University Press.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., … Latour, T. (2012a). The Genetics Lab: Acceptance and Psychometric Characteristics of a Computer-Based Microworld Assessing Complex Problem Solving. *Psychological Test and Assessment Modeling*, Vol. *54, No.* 1, pp. 54–72.

Sonnleitner, P., Brunner, M., Keller, U., Hazotte, C., Mayer, H., Latour, T., & Martin, R. (2012b). *The Genetics Lab_Theoretical background & psychometric evaluation (Research Report)*. Luxembourg, Luxemburg: University of Luxembourg.

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, Germany: Hogrefe.

Tapscott, D. (1998). *Growing up Digital: the Rise of the Net Generation*. New York: McGraw-Hill.

Terzis, V., & Economides, A. (2011). The acceptance and use of computer based assessment. *Computers & Education*, Vol. *56*, pp. 1032–1044.

Veen, W., & Vrakking, B. (2006). *Homo Zappiens - Growing up in a digital age*. London, UK: Network Continuum Education.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). Impact of goal specifity on strategy use and acquisition of problem structure. *Cognitive Science*, Vol. *20*, pp. 75–100.

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods*, Vol. *35, No.* 2, pp. 185–193.

Weiler, A. (2005). Information-seeking behavior in Generation Y students: Motivation, critical thinking, and learning theory. *The Journal of Academic Librarianship*, Vol. *31, No.* 1, pp. 46–53.

Winne, P. H. (2010). Improving Measurements of Self-Regulated Learning. *Educational Psychologist*, Vol. *45, No.* 4, pp. 267–276.

Wirth, J., & Klieme, E. (2003). Computer-based Assessment of Problem Solving Competence. *Assessment in Education: Principles, Policy & Practice*, Vol. *10, No.* 3, pp. 329–345.

Wood, R. T. A., Griffiths, M. D., Chappell, D., & Davies, M. N. O. (2004). The structural characteristics of video games: A psycho-structural analysis. *CyberPsychology & Behavior*, Vol. *7, No.* 1, pp. 1–10.

2.2. The Genetics Lab: Acceptance and Psychometric Characteristics of a Computer-Based Microworld Assessing Complex Problem Solving

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T.

(published in *Psychological Test and Assessment Modeling* 54 (2012), p. 54-72)

Psychological Test and Assessment Modeling, Volume 54, 2012 (1), 54-72

The *Genetics Lab*: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving

Philipp Sonnleitner[1], *Martin Brunner*[2], *Samuel Greiff*[3], *Joachim Funke*[3], *Ulrich Keller*[2], *Romain Martin*[2], *Cyril Hazotte*[4], *Hélène Mayer*[4] & *Thibaud Latour*[4]

Abstract

Computer-based problem solving scenarios or "microworlds" are contemporary assessment instruments frequently used to assess students' complex problem solving behavior – a key aspect of today's educational curricula and assessment frameworks. Surprisingly, almost nothing is known about their (1) acceptance or (2) psychometric characteristics in student populations. This article introduces the *Genetics Lab* (GL), a newly developed microworld, and addresses this lack of empirical data in two studies. Findings from Study 1, with a sample of 61 ninth graders, show that acceptance of the GL was high and that the internal consistencies of the scores obtained were satisfactory. In addition, meaningful intercorrelations between the scores supported the instrument's construct validity. Study 2 drew on data from 79 ninth graders in differing school types. Large to medium correlations with figural and numerical reasoning scores provided evidence for the instrument's construct validity. In terms of external validity, substantial correlations were found between academic performance and scores on the GL, most of which were higher than those observed between academic performance and the reasoning scales administered. In sum, this research closes an important empirical gap by (1) proving acceptance of the GL and (2) demonstrating satisfactory psychometric properties of its scores in student populations.

Key words: microworlds, complex problem solving, acceptance, computer-based testing, educational assessment

---

[1] *Correspondence concerning this article should be addressed to:* Philipp Sonnleitner, MSc, EMACS research unit, University of Luxembourg, Campus Walferdange, 7201 Walferdange, Luxembourg; email: philipp.sonnleitner@uni.lu

[2] EMACS research unit, University of Luxembourg

[3] Department of Psychology, University of Heidelberg

[4] Centre de Recherche Public Henri Tudor, Luxembourg-Kirchberg

Introduction

Many contemporary educational curricula and educational assessment frameworks (OECD, 2004, 2010) emphasize the critical importance of the (domain-general) ability to solve complex problems (e.g., Ridgway & McCusker, 2003) for occupational success and lifelong learning. Complex problem solving abilities are frequently assessed through so-called "microworlds," in which students solve problems in interactive, dynamic scenarios that capture both problem-solving processes and their products (Leutner, Funke, Klieme, & Wirth, 2005; Wirth & Funke, 2005).

In applied assessment, it is essential that the instruments administered are accepted by the test takers (and by those who use the scores obtained). For computer-based microworlds in particular, the acceptance concept may be meaningfully embedded in the theoretical framework of technology acceptance models (e.g. Terzis & Economides, 2011). These models distinguish several facets (e.g. perceived ease of use or attractivity) that contribute to test users' acceptance of an instrument.

Although it has been claimed that microworlds enjoy high acceptance among students because they use computer technology (Ridgway & McCusker, 2003), this assertion rests on the assumption that any computer-based instrument will meet the expectations of today's students. Yet these students are "digital natives" (Prensky, 2001), who expect software applications to demonstrate the highest quality in terms of usability, functioning, and design. Given the rapid pace of software development, microworlds are in constant need of being updated. However, the latest microworlds for which psychometric evaluations are available date back one (Kröner, 2001) or even more decades (Omodei & Wearing, 1995; Vollmeyer, Burns, & Holyoak, 1996). Moreover, to the best of our knowledge, the acceptance of these microworlds by student test takers has not yet been empirically investigated.

In addition, although complex problem solving (CPS) is an important competency to be acquired by all students, most previous studies on CPS have drawn on adult samples (e.g., psychology students), rather than on samples of school students. The few available studies with student samples (e.g., Kröner, 2001; Kröner, Plass, & Leutner, 2005; Rollett, 2008; Süß, 1996) have focused on students in the highest academic track, and usually at grade 10 or above.

Taken together, little is known about (1) the acceptance of (existing) microworlds among today's students or (2) whether the scores yielded by these microworlds are valid and reliable indicators of CPS of students in lower academic tracks or lower grade levels. Because we doubted that microworlds dating back to the last century would meet the expectations of today's students, we developed a new microworld: the *Genetics Lab* (GL). This article presents two studies examining the acceptance and psychometric properties of the GL in ninth grade students of the intermediate and highest academic track in Luxembourg.

Characteristics of the *Genetics Lab*

The GL is rooted in the so-called DYNAMIS framework, a widespread and established approach for the design of computer-based problem solving scenarios to study complex problem solving and decision making (cf. Funke, 1992, 1993, 2001). Within this framework, problem solving scenarios consist of several input variables (which can be manipulated by the test taker) and several output variables (which are connected to input and/or output variables via linear equations and cannot be directly manipulated). Scenarios in this tradition realize key characteristics of a complex problem in a standardized way as they can be described in terms of their complexity (number of variables), connectivity (number and type of the underlying connections), the degree of their "eigendynamic" (change of variables without intervention; see Frensch & Funke, 1995), intransparency (the underlying connections are hidden) and multiple goals (number of output variables which must be influenced). In order to distinguish between knowledge acquisition skills and knowledge application skills, working with such a scenario is divided into an "exploration" phase and a "control" phase.

In the present paper, we developed the new microworld GL using the DYNAMIS-framework because it allows for (a) clear and well-defined problem solutions, (b) the comparison between scenarios within a formal framework, (c) a separation of knowledge acquisition and knowledge application, and (d) the theoretically grounded derivation of scores to represent individuals' problem-solving performance in the exploration and control phase. Further, the GL also capitalizes on a current methodological advancement within the DYNAMIS tradition – the MicroDYN-approach (Greiff & Funke, 2010) – that combines problem-solving research grounded in experimental psychology with well-established principles from individual differences research and psychometrics (see also Süß, 1999). In particular, within the MicroDYN approach, test takers complete several scenarios of reduced complexity instead of one extensive scenario. Performance on these scenarios (like individual items of a performance scale) can be aggregated across scenarios to yield overall performance scores with considerably higher reliability than a single performance score obtained from one extensive scenario.

Task and performance scores

In the GL (Figure 1), the task of the students is to examine how the genes of fictitious creatures (input variables) influence their physical characteristics (output variables). In line with the DYNAMIS-approach, the examination of each creature is split into two consecutive phases: (a) the exploration phase and (b) the control phase.

In the exploration phase, students actively manipulate the creature's genes (Figure 1a). The effects of their genetic manipulations (i.e., user inputs) on characteristics (i.e., system outputs) are displayed in diagrams. By carefully analyzing this information, students learn about the underlying connections between genes and characteristics. As described above, the complexity of a creature depends on (a) the number of genes or characteristics, (b) the number of connections between them, (c) the kind of connection (positive or

negative), and (d) whether characteristics change without being affected by genes (eigendynamic).

Students' behavior while working on the GL is recorded in a detailed *log-file* which is used to derive performance scores as well as to validate whether students work properly on the GL (see below). Specifically, the log-file allows us to derive a process-oriented score reflecting how systematically students explored the creatures. Exploration is most informative for solving the task if students set one gene to "on" and all other genes to "off" – it is only then that changes in characteristics can be unambiguously attributed to the gene that is switched on (Vollmeyer et al., 1996). Moreover, eigendynamic is best detected by switching all genes off. The *Systematic Exploration* score indicates the average proportion of such informative steps to the total number of steps taken in the exploration phase across all creatures that were explored (Kröner et al., 2005).

At any time during the exploration phase, students can document their knowledge in a database (Figure 1b). We scored these records on the basis of an established scoring algorithm (see for example Funke, 1992, 1993 or Müller, 1993) that reflects knowledge about how a gene affects a certain characteristic of a creature and knowledge about the strength of such an effect. To this end, a student's knowledge about how genes affect the characteristics of a certain creature is compared to the true underlying relationships. Correctly identified relations yield higher knowledge scores. Note that these scores were corrected for guessing (i.e. an effect exists or does not, producing a guessing probability of .50 per effect) and weighted by the kind of knowledge. In line with previous studies, we emphasized relational knowledge by multiplying it with a weight of .75 whereas knowledge about the strength of an effect was weighted by .25 (Funke, 1992). Knowledge scores were derived for each creature in a first step, and then summed up across all creatures to compute a global *System Knowledge* score.

In the control phase, students are required to manipulate the genes to achieve specified target values on certain characteristics (Figure 1c). They are allowed to consult their records in the database during this phase. Note that these manipulations must be achieved within three steps, which forces students to plan their actions in advance – a key characteristic of successful problem solving (Funke, 2003). To score students' *Control Performance*, we applied a scoring algorithm based on the final deviations from the target values. For each creature, we computed the absolute difference between the specified target value and the achieved value for each affected output variable. This difference was then divided by the initial difference, thus taking into account whether and how strongly students succeed in reducing the difference between the starting values and the target values. The resulting ratios were summed up across creatures to derive a *Control Performance* score.

Advantages of the *Genetics Lab* relative to previous microworlds

Compared to previous microworlds, the GL has some features that may enhance the reliability and validity of the performance scores yielded. First, many previous microworlds were based on a single but very extensive problem scenario. This so-called one-item approach has severe shortcomings (Greiff & Funke, 2010; Kröner, 2001): (1)

when controlling the microworld, the test is "contra adaptive," as low performing test takers are confronted with situations of increasing difficulty – with every suboptimal control step, it becomes harder to achieve the goal values. (2) All performance indicators are merely based on the interaction of the test taker with one extensive item. Therefore, basic psychometric quality standards are violated. Simulation-based tests asking multiple-choice questions about different conditions of the system (e.g. Kröner, 2001; Kröner et al., 2005) do not solve this problem. There is still only one complex problem to be explored and controlled; the related items can be seen as an item-bundle "at best" (Greiff & Funke, 2010). As said above, the GL, in contrast, is based on the MicroDYN approach (Greiff & Funke, 2010), in which students examine several independent scenarios (i.e., several creatures). Students thus show their ability to deal with problems of varying complexity and content. As a consequence, aggregating performance scores across creatures yields more reliable scores of the students' ability to deal with complex problems than does a single scenario.

A second advantage of the GL over former microworlds is related to the fact that these have extensive written instructions or extensive training periods with varying levels of standardization (cf. Rollet, 2008). Both forms of instruction are somewhat problematic. First, when instructions are presented in the form of long texts, student performance in microworlds may be contaminated by their reading ability. Second, when training sessions are not highly standardized, student performance can hardly be compared across test administrations, since students may receive a different quantity and quality of learning opportunities. To overcome these problems, the instructions of our GL are based on standards for modern multimedia learning to ensure that students fully understood the task requirements (Mayer, 2005; Mayer & Moreno, 2003). After starting the GL, students work for about 15 minutes on automatized, interactive instructions which introduce each task of the GL (exploring the creature, drawing a causal model and achieving goal states) separately: After a short written explanation visualized by an animation, students may practice the specific task. For drawing the causal diagram and achieving the goal values, detailed visual feedback is provided. When questions arise during the exercises, students are directed to the built-in help function, which explains all symbols shown on the screen in written and visual form.

A third disadvantage of traditional microworlds overcome by the GL is their reliance on prior knowledge (e.g., Süß, 1996). The semantic embedding of the GL is entirely fictive, meaning that it makes very low demands on prior knowledge. A fourth disadvantage of previous microworlds not shared by the GL is their reliance on numerical input formats. This format renders the specific input values used critically important, as some input values make relationships much easier to detect than others, particularly when the scenario is based on linear equations. The GL, in contrast, uses an iconic input format (Figure 1). Thus,

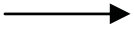
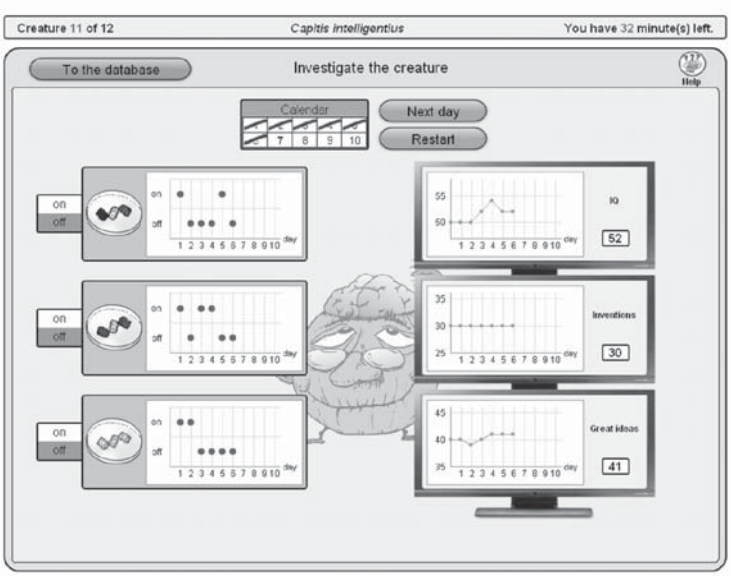
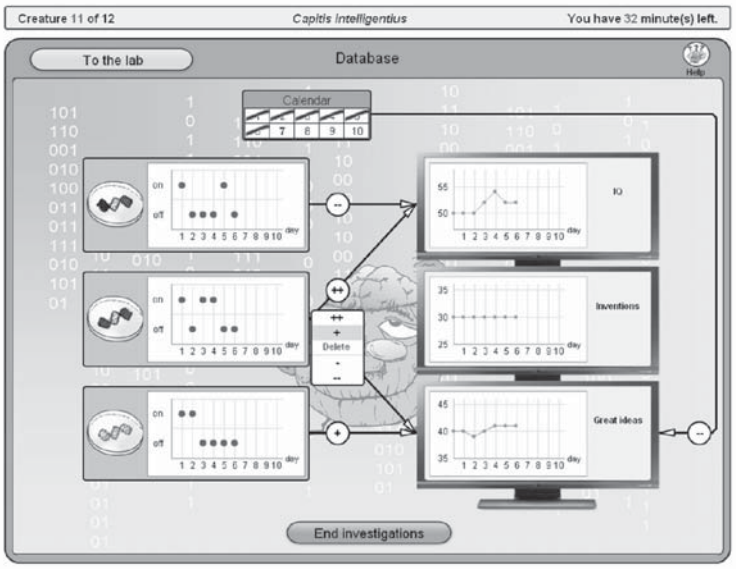


Figure 1: Screenshots of the different phases of the Genetics Lab: (a) Students explore how genes affect the characteristics of a fictitious creature and (b) record their knowledge in a database. (c) Students aim at achieving a given level of a characteristic (indicated by a red line and target value).



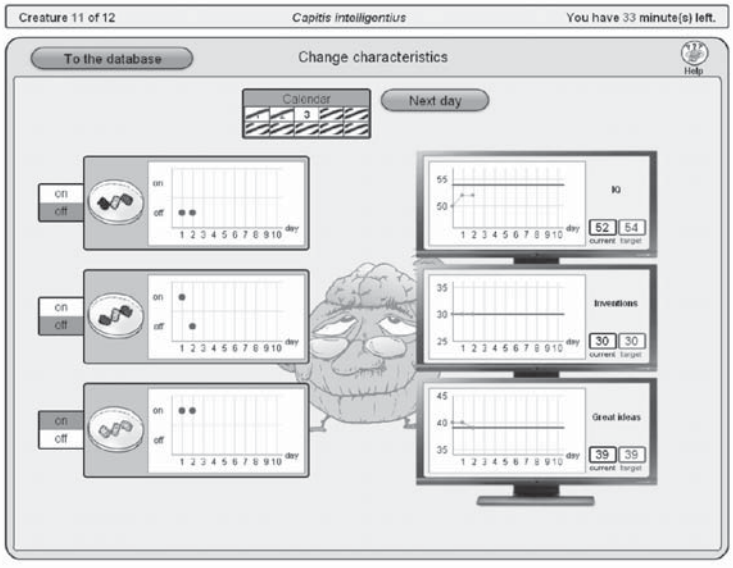
a. Phase 1: Exploring the creature

Students explore the effects of genes on certain characteristics of a number of organisms in a fictitious lab. By manipulating genes and observing the characteristics for a certain time, students can draw conclusions about the connections and formulate hypotheses that can then be tested.



b. Phase 1: Recording knowledge

Students document the knowledge they acquire about the relations between genes and characteristics in a database. Relations between genes and characteristics are expressed by means of arrows describing the type and strength of the connection. The resulting causal diagram can be interpreted as the theoretical model developed by the student exploring the creature.



c. Phase 2: Achieving target values

In the final phase, students have to manipulate the genes to alter the characteristics of organisms and reach specified target values. To this end, they can access the database in which they have recorded the knowledge previously acquired. This phase requires the competencies of using a theoretical model to inform concrete actions and controlling the resulting outcomes.

student scores are expected to be less dependent on arithmetic ability. A fifth advantage of the GL is its handling of "eigendynamic" effects. The interpretation of the scores yielded by previous microworlds including scenarios with "eigendynamic" was difficult, as high scores could be achieved by either high proficiency or by doing nothing (Kluge, 2008). The creatures in the GL are deliberately designed in such a way that all influences on characteristics are counterbalanced. Scores based on this "balanced" design have the advantage that they indicate whether (1) students actively explored the creature to detect eigendynamic(s), which are balanced out in the initial state, and whether (2) students took the eigendynamic into account in manipulating characteristics to achieve the specified target values. A sixth advantage of the GL relates to its attempt to increase test motivation and decrease test anxiety (McPherson & Burns, 2007; Washburn, 2003) by incorporating game-like characteristics (see Wood, Griffiths, Chappell, & Davies, 2004). These include immediate feedback in the form of scores reported after both phases have been completed for each creature, a semantic embedding of the scenario that puts the student into the role of a young scientist, and a comic-like design of the whole user interface (e.g., buttons and creatures) to ensure video-game like appearance. All of these features are aimed at eliciting maximum student performance.

Method

Aims and hypotheses

This article examines acceptance and psychometric properties of the GL in students. Specifically, in Study 1, we tested the hypothesis that the GL is accepted among students (Hypothesis 1). To our knowledge, this is the first time that user acceptance of a microworld has been investigated. Furthermore, Study 1 tested hypotheses relating to two important psychometric characteristics of the GL: (a) the construction rationale of the GL (e.g., multiple balanced scenarios, standardized instruction) yields reliable performance indicators of CPS (i.e. showing a high internal consistency) (Hypothesis 2); (b) meaningful intercorrelations of these scores provide preliminary evidence for their construct validity. In particular, in line with previous studies on CPS (e.g., Kröner, 2001; Kröner et al., 2005; Wirth & Funke, 2005), we expected S*ystematic Exploration* to have a positive influence on the *System Knowledge* acquired (Hypothesis 3), and *System Knowledge* to positively impact *Control Performance* (Hypothesis 4).

Study 2 aimed to replicate and significantly extend our psychometric evaluation of the GL. The version of the GL administered in this study contained fewer scenarios than that used in Study 1, thus making it possible to administer the test within a school lesson (a typical constraint on educational assessment). Nevertheless, we expected that this shorter version would yield comparably reliable scores (Hypothesis 5). We further examined the construct validity of the performance scores by analyzing three more hypotheses. Specifically, we expected to observe a similar pattern of score intercorrelations as in Study 1 (Hypothesis 6). Moreover, as the conceptual definition of intelligence focuses on reasoning and problem solving processes (e.g., Gottfredson, 1997), we expected – in line with

previous research (see Gonzales, Thomas, & Vanyukov, 2005; Kröner, 2001; Kröner et al., 2005; Rigas, Carling, & Brehmer, 2002; Wenke, Frensch, & Funke, 2005; Wirth & Funke, 2005) – to find a positive association between performance scores on the GL and intelligence measures (Hypothesis 7). Further, given the emphasis on CPS in educational curricula, we expected that GL performance scores would be positively correlated with academic performance as measured by grades (Hypothesis 8).

Data analysis

All data analyses were carried out using SPSS 11.5 for Windows. The type-I risk α for data analyses was set at $p < 0.05$, two-tailed.

Study 1

Participants and procedure

Participants in Study 1 were 61 ninth graders of an intermediate-track secondary school in Luxembourg. The school volunteered to participate in this study in order to explore the potential of the GL for use as an evaluation tool in science education. The study was conducted with approval from the Luxembourgish Ministry of Education and in accordance with the ethical standards of the University of Luxembourg and the data protection rules of the Luxembourgish commission for data protection (Commission nationale pour la protection des données). Both students and their parents were informed in written form about the scientific background of the study well in advance and were given the opportunity to refuse participation in the study.

Trained research assistants administered the GL and a questionnaire at school during regular class time. In addition, they observed the students' progress in working on the GL and pointed out the built-in help function if questions arose. To foster commitment, students were offered detailed written feedback on their performance after completion of the study. Nevertheless, data from 11 students were excluded because they did not work properly during the control phase (i.e., they skipped more than a quarter of the control phases). For (non-systematic) technical reasons, data from a further seven students had to be excluded. The final sample therefore comprised 43 students (19 females; $M = 15.8$ years; $SD = .87$ years). Note that Annex 1 presents the results as obtained for the student sample of Study 1 for whom complete data was available (i.e., $n = 54$ students).

Measures

Acceptance. We embedded our definition of acceptance in the conceptual framework of well-established technology acceptance models (e.g. Terzis & Economides, 2011). Within these models, the *Perceived Ease of Use* of an assessment instrument and its *Attractivity* are crucial factors that may contribute to its acceptance among potential

users. In addition, the *Comprehensibility* and *Functionality* of an assessment instrument are important factors determining its usability and thus its acceptance.

Consequently, students were asked to rate various elements of the GL (e.g., input format, help functions, diagrams; see Figure 1) on these four dimensions to help us investigate the GL's acceptance and usability among students and to identify any problems. The items used to assess these acceptance dimensions are listed in Annex 2. Students responded to these items on a 5-point rating scale with higher values indicating a more positive evaluation (see Note in Table 1 for a description of the verbal response anchors). Item scores were summarized to total scores indicating students' evaluation of each acceptance dimension. These total scores were expressed as a percentage of maximum possible scores that could be attained on a certain acceptance dimension (POMP, see Cohen, Cohen, Aiken, & West, 1999). In other words, a value of 0 indicates the lowest possible score, a value of 100 indicates the highest possible score, and values greater than 50 indicate that positive student evaluations outweigh negative evaluations on a certain acceptance dimension. Thus, we consider mean values above 50 % as positive outcomes. In addition, students stated whether they (a) had enjoyed working on the GL and (b) would like to complete the GL again (Yes/No). Given the lack of comparable studies or benchmarks, we see this approach as a reasonable way to get a balanced picture of the GL's acceptance.

Complex problem solving. The GL was administered without a time limit and contained 16 scenarios of varying complexity. Performance across scenarios was summarized by three scores indicating students' proficiency in (a) exploring the creatures (*Systematic Exploration*), (b) identifying the relationships between genes and the creatures' characteristics (*System Knowledge*), and (c) achieving specified target values on the creatures' characteristics (C*ontrol Performance*). These scores were (linearly) transformed into POMP scores with a value of 100 indicating the highest possible score.

Results and discussion

In terms of Hypothesis 1 concerning the acceptance of the GL (see Table 1), students rated the GL and its elements to be attractive (M = 64, SD = 22) and working with it to be fairly easy (M = 54, SD = 23). Moreover, 65 % of students reported that they enjoyed working on the test and 49 % that they would like to complete it again. Overall ratings of the GL's comprehensibility (M = 61, SD = 17) and functionality (M = 60, SD = 22) were also good. Close inspection of students' responses revealed that the instructions for the control phase were (particularly) hard to comprehend. This finding may explain the strong relationship between the *Control Performance* and *Acceptance* scales and why 11 students did not work properly during the control phase. In sum, these results indicate that the GL was generally accepted by students and thus support Hypothesis 1. Correlations with performance scores were positive, indicating that high-performing students accepted the GL more than low-performing students. Furthermore, the results on usability issues informed some improvements to the instructions that were made in Study 2.

Table 1: Means, standard deviations, reliability measures, and intercorrelations

	No. of items	α	M	SD	Min.	Max.	p25	MD	p75	SE	SK	CP	FI	NI	Math	Science
Study 1 (*n* = 43)																
Complex problem solving																
Systematic Exploration (SE)	16	.94	21	12	1	61	13	21	27	1			-[a]	-	-	-
System Knowledge (SK)	16	.89	54	12	38	96	46	51	57	**.54**	1		-	-	-	-
Control Performance (CP)	16	.80	74	13	36	99	67	72	85	**.40**	**.38**	1	-	-	-	-
Acceptance																
Perceived Ease of Use	4	.71	54	23	0	100	44	56	69	.31	**.44**	**.39**	-	-	-	-
Attractivity	9	.91	64	22	0	100	56	67	78	.22	**.34**	**.54**	-	-	-	-
Comprehensibility	10	.81	61	17	20	100	50	63	73	.28	**.49**	.29	-	-	-	-
Functionality	7	.82	60	22	0	100	46	64	71	.17	**.35**	**.56**	-	-	-	-
Study 2 (*n* = 61)																
Complex problem solving																
Systematic Exploration (SE)	12	.88	26	11	7	66	19	25	32	1						
System Knowledge (SK)	12	.77	53	12	35	100	45	51	59	**.35**	1					
Control Performance (CP)	12	.61	75	10	51	96	68	76	82	.24	**.47**	1				
Intelligence																
Figural Intelligence (FI)	20	.55	45	15	15	75	33	45	55	**.39**	**.40**	**.27**	1			
Numerical Intelligence (NI)	20	.88	71	23	5	100	60	70	90	.05	**.32**	**.34**	**.26**	1		
Academic performance																
Mathematics	-	-	68	19	25	100	57	67	83	**.39**	**.35**	**.37**	**.30**	.21	1	
Science	-	-	68	18	35	100	50	69	83	**.30**	.23	.16	**.29**	.15	**.65**	1

α = Cronbach's alpha.; p25 = first quartile (Q1); p75 = third quartile (Q3)

[a] FI, NI, and mathematics and science grades were not assessed in Study 1.

Note. All coefficients printed in bold are significant at $p < .05$ (2-sided significance).

In terms of the psychometric evaluation of the GL (Table 1), the performance scores showed satisfying levels of reliability (supporting Hypothesis 2). Internal consistencies (Cronbach's alpha α) ranged between α = .80 (*Control Performance*) and α = .94 (*Systematic Exploration*), indicating that students' problem solving behaviour was (relatively) consistent across scenarios. In line with previous studies (e.g., Kröner et al., 2005; Wirth & Funke, 2005), we found meaningful patterns of correlations among performance scores, pointing to their construct validity. Specifically, the more systematically a student explored a creature, the higher her or his *System Knowledge* (r = .54, p = .000) (supporting Hypothesis 3). Further, *System Knowledge* had a positive impact on *Control Performance* (r = .38, p = .011) (supporting Hypothesis 4). In sum, these results underscore the reliability of the performance scores yielded by the GL and provide initial evidence for their construct validity. Note that all results were fairly robust when those students who did not work properly during the control phase were also included for analyses. Detailed results including these students are shown in Annex 1. Importantly, means on all *Acceptance* scales still remain above 50 on the POMP-metric, indicating good acceptance of the GL in the (full) student sample.

Study 2

Participants and procedure

Participants in Study 2 were 79 ninth graders in intermediate- (n = 35) and academic-track secondary schools in Luxembourg. Recruiting arrangements paralleled those for Study 1. Unfortunately, data from 15 students again had to be excluded for (non-systematic) technical reasons. Data from a further 3 students were excluded because these students did not work properly during the control phase (i.e., they skipped more than a quarter of the control phases). The final sample therefore comprised 61 students (35 females; M = 15.5 years; SD = .61 years). Trained research assistants administered the testing material and students were again offered detailed written feedback in order to foster their commitment. Note that Annex 1 presents the results obtained from the student sample of Study 2 for which complete data was available (i.e., n = 64 students).

Measures

Complex problem solving. To allow administration of the GL within a school lesson (i.e., 50 minutes, of which 15 minutes were used for instruction), the GL was shortened to 12 scenarios. Further, the instructions (e.g., the explanation of the control phase) were modified slightly based on the results of Study 1. Scoring procedures paralleled those used in Study 1.

Intelligence and academic performance. Intelligence was measured by two subscales from the IST 2000 R, a widely used and well-elaborated German intelligence test (Amthauer, Brocke, Liepmann, & Beauducel, 2001). The *Selecting Figures* subscale is a measure of figural intelligence (FI); the *Number Completion* subscale is a measure of numerical intelligence (NI). Students' reports on their mathematics and science grades in

the last trimester were used as an indicator of *Academic Performance*. Both intelligence measures and grades were transformed into POMP scores.

Results and discussion

The internal consistency of all three performance scores was lower in Study 2 than in Study 1 (see Table 1), with values ranging from α = .61 (*Control Performance*) to α = .88 (*Systematic Exploration*). Thus, the results did not fully support Hypothesis 5. However, *Systematic Exploration* and *System Knowledge* showed acceptable reliability and the internal consistency of Control Performance may still be sufficient for research purposes – particularly when an assessment instrument is needed that can be administered during one school lesson.

Crucially, the GL performance scores showed the same pattern of intercorrelations as in Study 1 (supporting Hypothesis 6): *Systematic Exploration* again had a positive impact on *System Knowledge* (r = .35, p = .006), which in turn led to higher *Control Performance* (r = .47, p = .000). Our results also confirmed the conceptual relationship between CPS and intelligence (Hypothesis 7). Although the scale score measuring FI showed relatively low reliability and the scale measuring NI showed a ceiling effect, all GL performance scores were substantially related with these intelligence measures. Note that the strength of the relationship was comparable to that reported in previous studies (e.g., Kröner, 2001; Rigas et al., 2002). Further, we observed differential associations: FI was more strongly related to *Systematic Exploration* (r = .39, p = .002) and *System Knowledge* (r = .40, p = .001) than to *Control Performance* (r = .27, p = .035). One plausible explanation is that the exploration of creatures places strong demands on figural abilities (e.g., students need to interpret diagrams and to visualize their knowledge in the form of causal diagrams). NI was more strongly related to *System Knowledge* (r = .32, p = .011) and *Control Performance* (r = .34, p = .007); its relation to *Systematic Exploration* was negligible (r = .05, p = .729). One plausible explanation is that NI is required to determine the strength of an effect (yielding higher scores on *System Knowledge*) and to execute the computations needed to achieve the target values.

Finally, GL performance scores were positively related to both indicators of academic performance (supporting Hypothesis 8). However, we observed some differential relationships. Mathematics grade correlated positively with all performance scores, whereas science grade was more strongly related to *Systematic Exploration* than to the other two GL performance scores. Interestingly, grades tended to be more strongly associated with GL performance scores than were intelligence measures, for which a significant correlation with grades was to be expected (Gottfredson, 1997). Again, all results were fairly robust even when students who were identified as not properly working on the GL were included in the analyses (see Annex 1 for detailed results).

General discussion

Although reliable and valid assessment of CPS by means of microworlds has become increasingly important in the educational context, little is known about the psychometric characteristics of microworlds or their acceptance among students in lower academic tracks and grade levels. This article examined these questions in two samples of ninth graders in intermediate- and academic-track schools in Luxembourg who worked on the newly developed GL microworld. In developing the GL, we drew on (a) the DYNAMIS framework to conceptualize complex problem solving, (b) standards for modern multimedia learning (Mayer, 2005; Mayer & Moreno, 2003) and (c) game-like characteristics to increase test motivation and decrease test anxiety (Wood et al., 2004). Moreover, the GL also improves on previous microworlds by implementing relevant features like multiple balanced scenarios, standardized instructions, and iconic input format.

Today's students – most of whom are "digital natives" (Prensky, 2001) – expect software applications (e.g., video games) not only to demonstrate the highest quality in terms of usability and functioning, but also to be presented in an appealing design. Old-fashioned designs and cumbersome handling may therefore threaten the acceptance of computer-based tests. Our results showed that the GL was widely accepted among students. For example, *Perceived Ease of Use* and *Attractivity* – both common constructs in technology acceptance models (Terzis & Economides, 2011) – received high ratings. Moreover, when the GL's instructions were improved in Study 2, the number of students who skipped items – also a clear indicator of acceptance – decreased significantly. Thus, we provided initial empirical evidence that microworlds such as the GL can be applied in an educational context, where student acceptance is considered to be important.

Furthermore, both presented studies provided promising initial empirical evidence for the psychometric quality of the GL's performance indicators. First, in both studies, the GL's performance scores demonstrated high internal consistencies that were sufficient for research purposes. Note that the reliability of these scores can be enhanced by including more scenarios (e.g., when the GL is used for individual assessment). The construction of scenarios follows a pre-defined rationale and is therefore relatively easy and straightforward. Second, both studies provided initial evidence for the construct validity of these scores. In line with previous studies (e.g., Gonzales et al., 2005; Kröner, 2001; Kröner et al., 2005; Rigas et al., 2002), our findings confirmed a conceptual relationship between CPS and intelligence. The results suggested that the two phases of the GL are differentially affected by differing facets of inductive reasoning. Third, our findings show a strong relationship between the GL's performance scores and academic performance in terms of grades, which attests to the external validity of the GL and thus addresses the current lack of studies investigating the ability of microworlds to predict real-life criteria (Rigas et al., 2002). Moreover, this result underscores the importance of CPS in the educational context.

Despite the relatively large loss of data in both studies (18 data sets in each study), we doubt that the generalizability of our interpretations is affected. First, the loss of data caused by technical problems was non-systematic and therefore completely at random. In

Study 2, which investigated the GL's construct and external validity, this kind of data loss accounted for the vast majority of lost data sets (n = 15). Results on the GL's construct and external validity should therefore be robust against system-generated missing data.

Second, the article by Duckworth, Quinn, Lynam, Loeber, and Stouthamer-Loeber (2011) showed that test motivation may affect the validity of cognitive performance scores, particularly in research settings. The present study administered the GL in a low-stakes research setting where test motivation might have affected the results. One indicator for test motivation is the number of students who did not work properly on the GL: The number of students who were excluded because they skipped more than a quarter of the control phases was noteworthy in Study 1 (n = 11) but negligible in Study 2 (n = 3). Importantly, analyses including these students as shown in Annex 1 do not meaningfully differ from the analyses discussed above. Hence, these results suggest that the results on acceptance of the GL as well as on the psychometric properties of performance scores of the GL are not strongly biased when the analyses are based on a student sample where students differ in their motivation to work properly on the test as is to be expected in any low-stakes research situation.

Importantly, in identifying students who did not work properly on the GL we took full advantage of the possibilities of modern computer-based assessment by carefully studying students' log-files. This can be seen as a substantial advantage relative to traditional paper-pencil tests where such log-files do not exist. Using paper-pencil tests to identify such students is difficult, as this relies on strong theoretical assumptions about item response patterns or patterns of missing data.

In closing, despite promising initial empirical results on the acceptance of the GL and its psychometric properties, further studies are needed to replicate the findings of the present paper and to gain further insights into the psychometric properties of the GL (e.g., the factorial structure or measurement invariance across genders or students with differing migration backgrounds), and to elaborate on its validity in predicting real-life criteria. In order to promote this process, the GL will be published under an open-source license in English, French, and German during the first quarter of 2012. We look forward to its application to various research questions and different contexts.

Acknowledgements

This work was supported by funding from the National Research Fund Luxembourg (FNR/C08/LM/06). The authors would like to thank all the students and teachers for participating in this study and we would also like to thank Susannah Goss for the editorial support.

References

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

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Researc*h, 34 (3), 315-346. http://dx.doi.org/10.1207/S15327906MBR3403_2

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108* (19), 7716. http://dx.doi.org/10.1073/pnas.1018601108

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

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

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

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning, 7*(1), 69-89. http://dx.doi.org/10.1080/13546780042000046

Funke, J. (2003). *Problemlösendes Denken* [Problem-solving thinking]. Stuttgart, Germany: Kohlhammer.

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

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13-23. http://dx.doi.org/10.1016/S0160-2896(97)90011-8

Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme [Systematic investigation of complex problem solving using systems of minimal complexity]. *Zeitschrift für Pädagogische Psychologie, 56*, 216-227.

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

Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation* [Intelligence assessment via computer simulation]. Münster, Germany: Waxmann.

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., Funke, J., Klieme, E., & Wirth, J. (2005). Problemlösefähigkeit als fächerübergreifende Kompetenz [Problem solving as cross-curricular competence]. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 11-20). Wiesbaden, Germany: VS.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31-48). New York, NY: Cambridge University Press.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43-52. ttp://dx.doi.org/10.1207/S15326985EP3801_6

McPherson, J., & Burns, N. R. (2007). Gs Invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods, 39*, 876-883. http://dx.doi.org/10.3758/BF03192982

Müller, H. (1993). *Komplexes Problemlösen: Reliabilität und Wissen* [Complex Problem Solving: Reliability and knowledge]. Bonn, Germany: Holos

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

OECD. (2010). *PISA 2012 field trial problem solving framework. Draft subject to possible revision after the field trial.* Paris, France: OECD. Retrieved from http://www.pisa.oecd.org/dataoecd/8/42/46962005.pdf

Omodei, M. M., & Wearing, A. J. (1995). The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods, Instruments, & Computers, 27*, 303-316. http://dx.doi.org/10.3758/BF03200423

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon, 9,* 1-6. http://dx.doi.org/10.1108/10748120110424816

Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education, 10*(3), 309-328. http://dx.doi.org/10.1080/0969594032000148163

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

Rollett, W. (2008). *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme* [Strategy use, generated information and use of information in exploring and controlling complex, dynamic systems]. Berlin, Germany: Lit Verlag.

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, Germany: Hogrefe.

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

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer-based assessment. *Computers & Education, 56*, 1032-1044. http://dx.doi.org/10.1016/j.compedu.2010.11.017

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75-100. http://dx.doi.org/10.1207/s15516709cog2001_3

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, & Computers, 35*(2), 185-193. http://dx.doi.org/10.3758/BF03202541

Wenke, D., Frensch, P. A., & Funke, J. (2005). Complex problem solving and intelligence. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and intelligence* (pp. 160-187). New York, NY: Cambridge University Press.

Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme [Dynamic problem solving: Development and evaluation of a new assessment to control complex systems]. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55-72). Wiesbaden, Germany: VS.

Wood, R. T. A., Griffiths, M. D., Chappell, D., & Davies, M. N. O. (2004). The structural characteristics of video games: A psychostructural analysis. *Cyberpsychology & Behavior, 7*, 1-10. http://dx.doi.org/10.1089/109493104322820057

Annex 1: Means, standard deviations, reliability measures, and intercorrelations of samples including students who were identified as not working properly on the GL

	No. of items	α	M	SD	Min.	Max.	p25	MD	p75	SE	SK	CP	FI	NI	Math	Science
Study 1 (*n* = 54)																
Complex problem solving																
Systematic Exploration (SE)	16	.94	19	12	0	61	10	19	26	1			-[a]	-	-	-
System Knowledge (SK)	16	.89	52	11	38	96	46	49	55	**.56**	1		-	-	-	-
Control Performance (CP)	16	.84	70	15	33	99	62	69	79	**.48**	**.45**	1	-	-	-	-
Acceptance																
Perceived Ease of Use	4	.75	53	24	0	100	44	50	69	.21	**.38**	**.35**	-	-	-	-
Attractivity	9	.94	60	26	0	100	49	63	78	.25	**.35**	**.47**	-	-	-	-
Comprehensibility	10	.85	57	19	3	100	50	58	69	**.32**	**.45**	**.36**	-	-	-	-
Functionality	7	.86	59	24	0	100	46	64	72	.07	**.30**	**.38**	-	-	-	-
Study 2 (*n* = 64)																
Complex problem solving																
Systematic Exploration (SE)	12	.89	25	11	2	66	17	24	32	1						
System Knowledge (SK)	12	.76	53	12	35	100	45	50	59	**.34**	1					
Control Performance (CP)	12	.67	74	11	51	96	68	75	82	**.32**	**.43**	1				
Intelligence																
Figural Intelligence (FI)	20	.53	45	15	15	75	35	45	55	**.35**	**.40**	.20	1			
Numerical Intelligence (NI)	20	.87	71	23	5	100	60	70	90	.04	**.31**	**.31**	**.25**	1		
Academic performance																
Mathematics	-	-	68	18	25	100	56	67	83	**.39**	**.35**	**.36**	**.29**	.22	1	
Science	-	-	67	18	35	100	50	65	83	**.35**	.22	.25	**.26**	.14	**.63**	1

α = Cronbach's alpha.; p25 = first quartile (Q1); p75 = third quartile (Q3)

[a] FI, NI, and mathematics and science grades were not assessed in Study 1.

Note. All coefficients printed in bold are significant at $p < .05$ (2-sided significance).

Annex 2:
Items of the acceptance and usability questionnaire as applied in Study 1

Dimension (number of items; item stem) and corresponding items
Perceived Ease of Use (4 items; How easy were the following tasks for you?)[a]
(1) Investigating the fictive creatures, (2) Depicting my gathered knowledge, (3) Influencing the characteristics, (4) Reading the diagrams
Attractivity (8 items; How much did you like the following elements?)[b]
(1) Illustration of the creatures, (2) Design of the lab, (3) Design of the diagrams, (4) Feedback about your performance, (5) Using the database, (6) Influencing the characteristics, (7) Topic of a genetics lab, (8) Design of the test taken as a whole.
Comprehensibility (10 items; How would you rate the comprehensibility of the following aspects?)[c]
(1) Explanation of how the Genetics Lab works, (2) Explanation of how to depict your knowledge, (3) Explanation of how to influence the characteristics, (4) Help function, (5) Calendar, (6) Impact of time on the characteristics, (7) Different strengths of effects, (8) Layout of the diagrams, (9) Feedback about your performance, (10) Your task in general
Functionality (7 items; How well did the following elements work?)[d]
(1) Exercise at the beginning, (2) Switching the genes on and off, (3) Usage of the calendar, (4) Selection of the effect strengths, (5) Drawing the effects of genes, (6) Usage of the help function, (7) Confirmation of your depicted knowledge
Miscellaneous (2 items)[e]
(1) Did you enjoy working on the test?, (2) Would you like to repeat the test?

Note. Students used a five-point rating scale (labeled with 0,1,2,3, and 4) to evaluate the items of each dimension. The mimimum (i.e., 0) and maximum values (i.e., 4) were further labeled with a verbal anchor that varied across acceptance and usability dimensions, respectively.

[a]: Verbal anchors: very difficult (coded as 0) vs. very easy (coded as 4)

[b]: Verbal anchors: not at all (coded as 0) vs. very much (coded as 4)

[c]: Verbal anchors: incomprehensible (coded as 0) vs. very comprehensible (coded as 4)

[d]: Verbal anchors: did not work (coded as 0) vs. worked perfectly (coded as 4)

[e]: Answer options were yes and no

Chapter III – Validity and Fairness of the Genetics Lab

3.1. Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M.

(published in *Intelligence* 41 (2013), p. 289-305)

3.2. Differential relations between facets of complex problem solving and students' immigration background

Sonnleitner, P., Brunner, M., Keller, U., & Martin, R.

(published in *Journal of Educational Psychology* 106 (2014), p. 681-695)

Chapter IV – General Discussion

4.1. Summary of the main outcomes

To better deal with the cognitive heterogeneity of students, educational systems rely on valid and reliable instruments to capture these cognitive differences. Up to the present day, so-called intelligence tests fulfilled this purpose quite well, but were also criticized for not fully capturing students' cognitive abilities, only comprising static problems, and a lack of face validity. The present dissertation which is based on the COGSIM-project that was conducted from 2009 till 2012 at the University of Luxembourg, set out to determine the added value of complex problem solving scenarios for the assessment of students' cognitive abilities. In the course of the project, we first thoroughly developed and then psychometrically evaluated the complex problem-solving scenario Genetics Lab (GL).

4.1.1. Development of the Genetics Lab

The first study described in Chapter 2 (see 2.1.) showed that a considerable amount of today's students meet the characteristics of so-called "digital natives" or the "net generation" (Prensky, 2001; Tapscott, 1998), thus having high expectations concerning the design and handling of computer-based assessment tools. With the development of the GL we aimed to meet these expectations with an adapted test development process including extensive usability testing, the inclusion of multimedia components for explaining the task's demands, and the implementation of game-based characteristics to ensure test motivation, have proven to be successful in accommodating this delicate cohort. Compared to previous problem-solving scenarios, the GL has the advantage of using several independent scenarios of reduced

complexity to gather reliable performance scores (cf. Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015; Greiff, Wüstenberg, & Funke, 2012), a fictitious semantic embedding reducing the impact of previous knowledge to a minimum, and iconic input formats to make its handling more accessible.

Study 2 (see 2.2.) empirically showed that the GL enjoyed high acceptance among 9th grade students and in doing so, we were among the first to confirm the often anecdotal claim concerning the scenarios' good reception in the educational domain (e.g. Bennett, Jenkins, Persky, & Weiss, 2003; Ridgway & McCusker, 2003). Further, results of Study 2 also allowed for a satisfying psychometric evaluation of the Genetics Lab scale scores. In line with previous findings in the field, the three measured facets of complex problem solving behavior (i.e. gathering knowledge, documenting knowledge, applying knowledge; see 1.3.1.), were reliable and substantially associated with traditional measures of cognitive abilities, thus indicating construct validity. External validity of the GL's scores was supported by significant correlations with students' mathematics and science grades. Importantly, we were able to complement previous research on complex problem solving ability (CPS) by showing that established performance scores were also reliable and valid for students enrolled in lower grade levels (grade 9) and non-academic tracks.

Taken together, both publications support the notion that the development of the GL was successful. It enjoyed a high level of acceptance among a representative sample of 9th graders and initial results on its psychometric characteristics were promising. Given this important milestone, we next inspected the psychometric characteristics of the GL and its potential benefits in assessing students' cognitive abilities more closely.

4.1.2. Psychometric structure, external validity and potential added value of the Genetics Lab

In Chapter 3, we report further evidence concerning the GL's psychometric quality. Study 3 (see 3.1.) was based on a large-scale study including a representative number of 563 Luxembourgish students of different school tracks (academic and non-academic) and grade levels (9 and 11). Its major aim was a thorough investigation of the GL's underlying psychometric structure, its construct validity by studying the relation to reasoning ability, and the GL's capability to explain students' educational success. Given different theoretical accounts of the latent structure of CPS, ranging from a one-dimensional, general CPS ability (e.g. Abele et al., 2012) to a faceted conceptualization, distinguishing between two or three distinct facets of CPS (e.g., Greiff et al., 2012; Wüstenberg, Greiff, & Funke, 2012), we took a balanced stance, and investigated several psychometric conceptualizations of CPS. Whereas a one-dimensional conceptualization of CPS showed insufficient model fit, the data supported a faceted model of CPS, including the abilities to explore a problem, gather knowledge, and apply it accordingly to achieve given targets. Note that accounting for the high correlations between these facets within a hierarchical measurement model, including a general CPS ability factor at the top of the apex, was empirically equally well supported. Consequently, given the inconclusive theoretical stance on the best psychometric representation of CPS and the empirical plausibility of a faceted, as well as a hierarchical conceptualization, we employed both measurement models when studying the relation between CPS and reasoning ability.

A juxtaposition of psychometric conceptualizations (hierarchical vs. faceted measurement models) for both constructs confirmed that reasoning ability plays an important part in solving complex problems with correlations ranging from .38 between reasoning and the ability to explore a problem, to .59 between reasoning and knowledge application. The general CPS factor was also substantially associated with reasoning (r = .62). Despite this

strong connection, though, results supported the notion of CPS being an independent construct. When we accounted for the hierarchy of cognitive abilities in two nested factor models that acknowledged the influence of a general cognitive ability factor *g* on reasoning ability as well as CPS, the psychometric structure of the GL, including three distinct facets of CPS, remained stable. Thus, CPS seems to be closely related to reasoning but nevertheless a distinct construct.

After investigating the psychometric structure of CPS and its relation to reasoning, we turned to the idea of the potential benefits of problem-solving scenarios. Results mainly showed that the (impressive) external validity of facets of CPS, as well as the general CPS factor, is mostly attributable to the common variance core that is shared with reasoning. After controlling for *g*, (latent) correlations between CPS and its facets with different levels of educational success (specific subjects vs. general educational success) in most cases dropped to insignificant levels. Thus, a preliminary conclusion concerning the added value of problem-solving scenarios in the educational context pointed more to the (new) way problem-solving processes are captured than to the claim that CPS captures something more than traditional tests of cognitive abilities do (e.g., Funke, 2010; Greiff et al., 2012; Wüstenberg et al., 2012).

Study 4 (see 3.2.) directly aimed at the exploration of the potential benefits of complex problem-solving scenarios within the Luxembourgish context. It has previously been argued that tracking decisions in Luxembourg rely heavily on students' language skills and is therefore disadvantageous to immigrant students, who are more likely to be enrolled in lower school tracks than would be suited for their cognitive potential (e.g. Burton & Martin, 2008; Klapproth, Glock, Böhmer, Krolak-Schwerdt, & Martin, 2012). Since traditional tests of cognitive abilities were found to be influenced by students' educational background (Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Gustafsson, 2008), such measures probably merely confirm previous (inadequate) tracking decisions when evaluating students' cognitive potential to decide whether they should pursue an academic or a vocational career. Thus, we

studied the benefit of complex problem-solving scenarios in this context whose task demands are supposedly less trained for in school and which incorporate several characteristics that could minimize the impact of language (e.g., possibility to choose the preferred test language, mostly language-free and multi-media instructions).

Initially, we were able to show that the scales of the GL are indeed measurement-invariant, and thus fair with regard to students' immigration backgrounds, regardless of whether CPS was represented as faceted construct or as a general CPS ability. In doing so, we substantially contributed to previous research that showed promising results concerning measurement invariance of such scenarios with regard to other criteria, such as sex or nationality (Greiff et al., 2013; Wüstenberg, Greiff, Molnár, & Funke, 2014), but at the same time reported culturally different exploration strategies (Güss, Tuason, & Gerhard, 2009; Strohschneider & Güss, 1999). Thus, we confirmed that the GL allowed for fair comparisons between immigrant students and their native peers.

In line with previous studies in other domains (e.g., Martin, Liem, Mok, & Xu, 2012; Schleicher, 2006), however, we found that immigrant students were generally outperformed by their native peers in reasoning ability as well as CPS. Crucially, adopting the even-handed approach from the previous study, and conceptualizing CPS as either hierarchical or as faceted construct, revealed that immigrant students showed, in fact, better performance than their peers when exploring the problem scenarios but lost this advantage in each consecutive step. Thus, immigrant students are capable of systematically exploring a problem, but are less proficient in translating the generated information into declarative knowledge or even to applying it to achieve certain target states. Taking into account that most immigrant students of our sample were enrolled in the lower academic track in two multiple-indicator, multiple-cause models (MIMIC; Jöreskog & Goldberger, 1975; Muthén, 1989), it turned out that performance differences in reasoning, as well as CPS, were largely explained by the groups' differing

educational background. Results also showed that performance in the GL was less influenced by the attended academic track than were the administered reasoning scales and that immigrant students showed a more efficient exploration strategy despite their educational background. Although immigrant students' advantage in problem exploration was only small on a manifest (scale) level, this advantage was more pronounced in the lower academic track, supporting the notion of problem- solving scenarios as probably being a better-suited indicator of such students' cognitive potential than traditional tools used for this purpose (e.g., reasoning scales).

In sum, in Studies 3 and 4, we could show that the Genetics Lab reliably captures three distinct facets of CPS and displays satisfying construct and external validity. Although the Genetics Lab might not explain incremental variance in external criteria, such as educational success, compared to conventional tests of cognitive abilities, the differing approach to measuring problem-solving processes that are also present in (conventional) reasoning tasks might be the real benefit.

4.2. Theoretical and practical implications

The empirical results, as well as the methodological solutions, that originated from the present dissertation and its applied setting imply a number of theoretical and practical implications that can be grouped and shall be discussed in the following themes. After systematically evaluating the potential for complex problem-solving scenarios for the assessment of students' cognitive abilities on the basis of the conducted studies, we further discuss the results' implications on the interpretation of students' performances in these scenarios or, in other words, the validity of complex problem solving itself. Finally, we address

new challenges but also promising avenues for the domain of psychological assessment in general.

4.2.1. The potential of CPS scenarios for the assessment of students‘ cognitive abilities

The present dissertation set out to determine the potential of using problem-solving scenarios to assess students' cognitive abilities. Usually, this is done with conventional, paper-and-pencil-based intelligence tests, which were criticized in several aspects. Thus, before evaluating the added value of such scenarios for different assessment contexts in the educational setting, we first discuss to what extent they can answer the limitations of traditional measures of cognitive abilities.

One major critique concerned the coverage of the latent construct that those tests should measure (see 1.2.3.). Tests of cognitive ability were repeatedly criticized as not adequately capturing the facets of learning and problem solving (Dörner & Kreuzig, 1983; Snyderman & Rothman, 1987), which were seen as essential components of general cognitive ability (e.g. Gottfredson, 1997). Due to their computer-based administration, however, complex problem-solving scenarios might offer a valuable alternative in this context. In Studies 2 and 3, it was unanimously shown that the GL reliably captures different phases of the problem-solving process. Whereas traditional tests of cognitive abilities only capture the final "product" of the problem-solving process by scoring whether the solution to a problem has been found or not, the GL would allow for tracing where the problem-solving process has failed. For example, results of Study 4 showed that students with immigration background explored problems more efficiently than did their native peers, but lost this advantage in every consecutive step of the problem-solving process. Results also showed that the administered reasoning scale did not capture this advantage. In terms of measuring the problem-solving process, the GL thus clearly

provides benefits compared to traditional tests by differentiating between problem-solving strategies, gathered knowledge about the problem, and the ability to apply this knowledge in order to solve the problem. Note that this transparent and comprehensible way of capturing the whole problem-solving process also leads to a high face validity for problem-solving scenarios, an attribute that was significantly missed in traditional tests of cognitive abilities.

One might argue that the typical distinction of conventional tests between verbal, numerical, and figural problem-solving ability would be lost and that the GL's performance scores might be a hard-to-interpret conglomerate of these distinct reasoning skills, as was shown in the correlational pattern of Study 2 (see 2.2.). However, in terms of predicting or understanding real-life problem-solving ability, such a conglomerate, split in distinct problem-solving phases, might be more fruitful, since such problems are most often also not purely verbal, numerical, or figural in nature. Regarding problem-solving as a process would also allow for interventional programs targeting specific phases. It also has to be stated that the theoretical conceptualization of problem scenarios by clear (problem) characteristics (see 1.3.), per se, is superior to the (static, transparent, and well-structured) problems used in traditional cognitive ability tests.

Concerning the facet of learning, such scenarios could extend the mere memorizing of traditional tests by helping researchers to better understand why some students "learn" more out of the generated information during the exploration phase than others, how this information is translated into declarative knowledge, and how declarative knowledge, in turn, is different from, and interacts with, procedural knowledge that is captured in the final phase of problem-solving scenarios. Those interpretations, however, strongly rely on the assumption that the consecutive problem-solving processes causally influence each other. This will be critically discussed later in the chapter (see 4.2.2.).

Together with results of Studies 2 and 3, that confirm the previously established strong role of reasoning ability (another central facet of GCA; see Gottfredson, 1997) in problem-solving performance, we may conclude that problem-solving scenarios are a promising alternative to conventional tests for more adequately capturing GCA. The downside of these scenarios, however, lies in the measurement of GCA as a compound of abilities, not differentiating between the various facets of GCA but allowing for studying their concerted functioning throughout different phases of the problem solving process.

In the following, we subsequently discuss the potential advantages of this different approach to capture GCA for the typical educational contexts in which cognitive ability tests are used. Although the conducted studies do not cover all of these typical contexts, the results at least allow for speculating about the problem-solving scenarios' potential for different educational settings. A common solution for dealing with the cognitive heterogeneity of children is to place students in ability-dependent school tracks (Hallinan, 1994; Worthen, White, Fan, & Sudweeks, 1999). Due to problems with the conventional criteria on which such placement decisions are based (e.g., the influence of parents' socio-economic status, or gender and immigration background of the student), it has been argued that the use of objective measures of cognitive ability enable a more valid decision (cf. Heller, 1991; Maaz, Baeriswyl, & Trautwein, 2013).

Most of these decisions are made directly after elementary school, and traditional tests of cognitive abilities have proven to be of great value in this context (Anastasi & Urbina, 1997). Given that intelligence tests were initially developed for the purposes of predicting success in school (see 1.2.), this is not surprising. So, how could the use of problem-solving scenarios complement conventional tests in this context? Studies 2 and 3 clearly showed that the

performance indicators of CPS are substantially linked to educational success (see 2.2. and 3.1.). Although the results were, strictly speaking, only postdictive in explaining past achievements (grades, PISA & ÉpStan-scores), supposing that a reasonable stability of cognitive abilities at this age (Schalke et al., 2013), a prediction of educational success should also be possible. However, problem-solving scenarios were mainly studied and calibrated with university student samples (see 1.3.), and the Genetics Lab was mainly developed for students in grades 9 and above, thus already touching the youngest sample for which experiences with such scenarios exist.

Although meanwhile complex problem-solving scenarios were successfully applied within PISA (e.g. Greiff et al., 2014; OECD, 2014) and our studies, showing that it is possible to gather reliable and psychometric sound scores within this age group, experiences during test administration and a look at the performance distributions in Studies 2 and 3 clearly indicate that the scenarios are already quite demanding at this age. Making them less complex in order to also accommodate younger samples would theoretically be possible, but the scenarios would lose crucial characteristics, for example dynamically changing variables. Given the importance of this specific characteristic for the identification of the CPS facet problem exploration (see Study 3), however, this would barely make sense. A possible solution to this problem could be the contextualization of the applied problem scenarios. For example, Cosmides and Tooby (1992) could show that using a more natural context for a relatively abstract reasoning task (the Wason card selection task; Wason, 1968) significantly increased the performance of the test takers. Consequently, using a more natural semantic embedding for problem scenarios could make them also more accessible for younger or less able samples. On the other hand, it should then be taken care that the impact of previous knowledge is kept to an absolute minimum. Only future studies could show if such an approach would be feasible.

Thus, although problem-solving scenarios are not (yet) suited for tracking decisions right after elementary school, they could be an interesting alternative for tracking decisions within secondary school. For example, in Luxembourg, after grade 10, students can specialize within the secondary school tracks they are in. In addition, it is also possible to switch between tracks, if a student constantly performs at a higher or lower level than the track demands. Study 4, impressively shows the potential of problem solving scenarios to detect cognitive "underachievers" among immigrant students, and highlights that such scenarios are less influenced by previous educational experiences than are traditional tests of cognitive abilities. Thus, tracking decisions within secondary school could be made fairer and would substantially benefit from the use of problem solving scenarios as a measure of students' cognitive potential. Note, however, that the advantage of being less influenced by previous educational treatment could vanish, as soon as those scenarios become also part of the curriculum (e.g. Ridgway & McCusker, 2003).

A second purpose of cognitive ability tests within education is the identification of gifted or slow learners (Anastasi & Urbina, 1997; Worthen et al., 1999). For both groups, a complementary administration of problem-solving scenarios could provide advantages. In the first case, the (theoretically possible) very high complexity of problem-solving scenarios could be a real advantage. Conventional tests of cognitive abilities are designed and calibrated for students of average ability and special tests have to be used in order to diagnose "giftedness" (e.g. Preckel & Baudson, 2013; Shavinina, 2009). Characteristics of problem-solving scenarios, however, could relatively easily be altered so as to be very demanding even for gifted students, without leaving the theoretical rationale of scenarios used for average-skilled students. A comparison on the same underlying construct may thus be easier than when a totally different test is used. In addition, the differentiation of distinct problem solving phases may

allow for a better understanding of what separates gifted from “normal” students when they engage with problems, possibly leading to training programs for the latter.

The disentanglement of the problem-solving process may also be of use in understanding the origins of learning difficulties. The finding of Study 4, implying that immigrant students have specific difficulties in translating generated information into declarative knowledge, and hence use it, points to the potential of such scenarios to more precisely detect the problems students with learning difficulties might have. On the basis of such an evaluation, targeted interventions could take place – for example, helping students to develop strategies to more efficiently interpret given information and build knowledge out of it. Still, the application of problem-solving scenarios in both contexts has still to be explored to allow for a sound judgment of their benefits. Nevertheless, from a theoretical point of view, such scenarios comprise a big potential for a better understanding of giftedness and also learning disabilities.

Similar to supporting tracking decisions within secondary school, problem-solving scenarios could also be useful in (school) career counseling. As outlined in 1.1.3, career decisions are often guided by students’ specific cognitive strengths and weaknesses (cf. (Anastasi & Urbina, 1997; Heller, 1991; Worthen et al., 1999). For finding the best-fitting vocational or educational environment, however, a reliable measurement of students’ cognitive abilities is necessary. Again, as Study 4 showed, an advantage of problem-solving scenarios in this context could be a more valid view on, for example, immigrant students’ cognitive potential. On the other hand, results on problem-solving capabilities could also be the base for identifying fields of development, thus pointing to specific trainings going along with the chosen educational or vocational formation.

A final field of cognitive ability testing within the educational field, identified in 1.1.4, refers to an increased demand of teaching problem solving or even GCA within school curricula (Adey, Csapó, Demetriou, Hautamäki, & Shayer, 2007; Becker et al., 2012; Greiff et al., 2014; Kuhn, 2009; Martinez, 2000, 2013; Mayer & Wittrock, 1996). Related cognitive assessment would either be formative, evaluating students' progress throughout the course of the curriculum, or summative, evaluating the specific curriculum's general efficiency. By proving the psychometric soundness of the GL, we have shown that problem-solving scenarios validly capture students' problem-solving capacity and thus, are generally suited for this purpose. Several reasons may render such scenarios superior to traditional tests of cognitive abilities in this context. First, due to their conceptualization, they seem to better represent the GCA facet of problem solving. Even if largely similar cognitive processes are measured like in traditional reasoning tasks (see results of Study 3), the differentiation between distinct facets of problem solving has also the advantage of being more open to interventions and specific trainings. As discussed above, identifying specific weaknesses like immigrant students' difficulty to adequately use information they generate while exploring a problem, could be the starting point of a related intervention. A final advantage of problem-solving scenarios in this context is that they enjoy much higher acceptance among educators (e.g. Ridgway & McCusker, 2003), so that they were even included in widely acknowledged large-scale assessments like PISA (cf. OECD, 2010, 2014).

Thus, it seems that with such scenarios, it is socially more acceptable to measure students' cognitive potential than with traditional tests. Such scenarios reliably measure students' cognitive abilities but without explicitly touching the highly controversial term of "intelligence" (e.g. Adey et al., 2007). Given the importance of an adequate cognitive assessment within education (Adey et al., 2007; Becker et al., 2012; Martinez, 2000, 2013), however, being embraced by educators and policy makers in this domain is a crucial advantage.

Problem-solving scenarios could thus give guidance concerning the way to train cognitive abilities within curricula and open up possibilities to conduct better research on the interplay between GCA and education itself. Studying possibilities to make the scenarios also more accessible for less able students, as was discussed above, would be of high importance in this context, especially in the light of increasing efforts to guarantee an inclusion of students with special educational needs (cf. UNESCO, 2009).

4.2.2. The interpretation of students' complex problem-solving performances

Several implications arise concerning an adequate interpretation and conclusion about what causes students' performances in problem-solving scenarios. In other words, we tackle the issue of validity (cf. Borsboom, Mellenbergh, & Van Heerden, 2004). The psychometric analysis of the GL's underlying structure in Study 3 (see 3.1.) brought up evidence that CPS can be thought of as either a faceted construct, including three distinct facets of CPS, or as a hierarchical construct, including the three facets but positing an additional general CPS ability factor that influences each facet (3.1., Fig. 2). Given the continual lack of widely-acknowledged theoretical models and definitions of CPS (Fischer, Greiff, & Funke, 2011; Frensch & Funke, 1995; Quesada, Kintsch, & Gomez, 2005), neither of these conceptualizations is superior. However, each measurement model has its own implications on how to interpret students' performance.

The faceted model of CPS mirrors the consecutive phases of the Genetics Lab and includes the ability to (a) efficiently explore the problem scenario; (b) represent the gathered knowledge within a causal diagram; and (c) to use this knowledge to achieve certain target states. Although a commonly accepted definition of CPS is still missing, wide-spread consensus exists concerning its measurement (Frensch & Funke, 1995; Funke, 2003, 2010).

Thus, the applied operationalization of all three facets of CPS (see 2.1.4. for a detailed description) seems to be well grounded. However, in the light of previous studies, we have to be cautious when drawing direct inferences to underlying latent variables. The studies by Greiff et al. (2012) and Wüstenberg et al. (2012) both drew on a similar methodological approach as the GL by using several problem solving scenarios of reduced complexity. Crucially though, whereas Greiff et al. (2012) could support previous (methodologically different) findings of three distinct CPS-facets (Kröner, 2001), Wüstenberg et al. (2012) were only able to distinguish between the ability to depict gathered knowledge and the ability to reach target states. Consequently, as already noted in 3.1, whether the third CPS-facet of problem exploration can be identified as distinct factor, seems to strongly depend on surface, as well as content characteristics, of the used problem-solving scenario, thus rendering a clear inference of what is measured by the scenario difficult.

The use of dynamically changing variables, as well as operationalizing problem exploration ability via the applied strategy's efficiency as is done in the GL, seems to allow for the identification of this third facet. Whether the thorough development of the GL, including the comparably comprehensive instruction phase, contributes to this fact remains open to debate. Crucially, however, these inconclusive findings point to the necessity of a closer inspection of the circumstances where the facet of problem exploration can be identified. Reducing the assessment of CPS to two facets (knowledge and application) and ignoring problem exploration, as is done in some studies on CPS (e.g. Greiff et al., 2013), or focusing only on general CPS like in the PISA study (OECD, 2010), is probably a suboptimal way, since it ignores the promising potential this facet offers for the assessment of cognitive abilities as was shown in Study 4.

Another important aspect in the context of drawing conclusions about the scores' underlying constructs lies in the crucial role reasoning ability plays in problem-solving

performance. In line with previous studies, Study 3 showed that reasoning ability is strongly associated with all three facets of CPS, ranging from .38 to .59. Theoretically, this is explained by the notion that problem-solving scenarios capture basic, as well as higher-order, thinking processes (e.g., Funke, 2010). Such "basic" thinking processes are also measured with conventional reasoning tests causing the substantial correlations.

There seems to be empirical support that in capturing such higher-order thinking processes, problem-solving scenarios possess incremental validity over and above traditional measures of reasoning (Danner, Hagemann, Schankin, Hager, & Funke, 2011; Greiff, Fischer, et al., 2013; Greiff, Wüstenberg, et al., 2013; Wüstenberg et al., 2012). However, in Study 3 – contrary to other studies that account for the hierarchy of cognitive abilities – it turned out that only the common variance core that is shared between reasoning and facets of CPS could explain variance in external criteria (see 3.1., Table 3). Although the faceted structure of CPS remained stable, the remaining facet-specific variance was only associated with performance in the computer-based ÉpStan scores. In 3.1., we speculated that this might be attributable to the shared mode of test administration (i.e., computer-based) or to a lack of suitable external criteria. Together with the CPS psychometric structure's dependence on characteristics of the applied problem-solving scenario (see above); however, the question arises that if there is a strong empirical case that supports the interpretation of the reasoning/specific CPS facet – conglomerates as real, existing, latent entities – as a precondition for the performance scores' validity (see Borsboom, Mellenbergh, & Van Heerden, 2003; Edwards & Bagozzi, 2000; Markus & Borsboom, 2013). Put differently, results are inconclusive as to whether performance (for example, in problem exploration) is caused by underlying (latent) problem exploration ability.

In the light of the weak theoretical fundamentals of CPS and the strong empirical support and robustness of its indicators (i.e., the widely established scoring procedures), one

might even speculate whether a formative measurement model might be more suitable until there is more evidence on CPS's psychometric structure. Contrary to reflective measurement models (as applied in Studies 3 and 4 and most commonly in psychology) that assume a latent variable causes variance in the related indicators, formative measurement models incorporate the idea that a latent variable is more like a summary of the indicators used (Borsboom et al., 2003; Edwards & Bagozzi, 2000). A popular and widely-cited example for a formative model is the measurement of socioeconomic status (SES). SES is formed by, and summarizes, several indicators like income or educational level. Adopting this line of reasoning, facets of CPS would be a summary of how well students performed in each problem-solving phase, without indicating that problem exploration, knowledge declaration, and knowledge application exist as distinct, latent variables. CPS would then be defined through the scenario's characteristics (e.g., the competence to deal with complex, dynamic, interconnected, and opaque problems) without a specific claim concerning a latent ability that exists independently of the problem-solving scenarios. Note that CPS, when thought of as an index variable, would be equally useful in the educational setting, and still be a reliable indicator of students' problem- solving competence but without making (empirically not yet justified) assumptions about the underlying processes.

Another somewhat problematic "causal" explanation within the faceted model can be found in the temporal sequence of the problem-solving phases (*explore*, *depict knowledge*, and *apply knowledge*). This leads to the reasonable conclusion that the more efficient the applied exploration strategy, the higher the knowledge about the problem's structure, and the higher the performance in reaching the goals in the final phase. It seems as if this interpretation is also supported by the pattern of (latent) correlations between the three facets (see 3.1., Figure 2). Exploration efficiency is closely related to declared knowledge (r = .68) than to target achievement (r = .63), whereas knowledge shows the highest correlation with the ability to

achieve goals (r = .94). Note that this stance is also adopted when interpreting performance differences between immigrant students and their native peers in Study 4 (see 3.2.).

However, the observed correlational pattern emerged in a between-subject measurement model with summarized scenario performances and cannot directly be translated into within-subject cognitive processes within a single problem-solving scenario (Borsboom et al., 2003; Edwards & Bagozzi, 2000; Markus & Borsboom, 2013). Although the interpretation that the more efficiently a student explores a problem, the more knowledge he will acquire about it seems, strictly speaking, to be obvious, it is not supported by the applied (between-subjects) measurement model. Support for such an interpretation could, for example, be gained through a detailed analysis of how each student performs throughout the (problem-solving) phases in several scenarios or on several occasions. Such time series data would allow for unanimously understanding causation between the facets of CPS on an individual level. A different approach would be an experimentally controlled training of students in specific facets followed by the study of the training's effects on the different facets (cf. Markus & Borsboom, 2013).

Turning to the hierarchical measurement model of CPS, problems of interpretation remain. In Study 3, it turned out that the strong associations between CPS's facets could be modeled by a hierarchically superior general CPS ability factor that influences the facets. However, similar to the interpretational problems concerning the general cognitive ability factor *g* in intelligence research (Hunt, 2011; Van Der Maas et al., 2006), it is unclear what general CPS actually is and means. Importantly, controlling for general cognitive ability in Model *F* (hierarchical intelligence – faceted CPS; see 3.1, Fig. 3) did not influence the intercorrelation between CPS's facets and hence, not account for this general CPS factor. Thus, the facets of CPS share variance over and above the common reasoning processes.

But what could cause this common variance? One possible explanation could be inspired by alternative explanations of *g* in contemporary intelligence research. In their mutualism model, van der Maas and colleagues (2006) suggested that *g* arises through interactions between first-order factors of GCA. Accordingly, throughout cognitive development, high verbal ability would be beneficial for acquiring spatial ability, which in turn could improve the ability to deal with numbers and quantities. These mutual influences would result in a statistical *g*-factor, without corresponding to a real, existing, latent factor. Although the plausibility of this model has thus far only been supported by simulation, it is excellent food for thought, and also for interpreting the general CPS factor. In this line of thinking, it could represent the mutual influences of the separate facets of CPS. Exploration behavior influences knowledge about the problem, which in turn influences the ability to systematically manipulate the problem. However, there are problems associated with a causal interpretation of the correlations (see above). Another possible explanation could be that the general CPS ability factor illustrates a mere method factor, which captures students' ability to deal with computer-based assessment instruments. Similar to the facets of CPS, after controlling for the general cognitive ability factor in Model *K* (hierarchical intelligence – hierarchical CPS, see 3.1., Table 3), the specific variance of general CPS was only substantially correlated with the computer-based ÉpStan-scores.

Given the still-uncertain situation in which it is uncertain what general CPS ability could represent, a preliminary conclusion could again be an adapted measurement model to account for this lack of theoretical foundation. Even if the captured facets of CPS are thought of as real, existing, distinct, latent variables (see above), the upper part of the measurement model including general CPS could be formulated as a formative measurement model. The result would be a so-called mixed measurement model, in which the higher-order construct is a formative index of the lower-order reflective latent constructs (p.122, Markus & Borsboom,

2013). General CPS would then be more of a proxy for indicating students' overall performance in problem-solving scenarios. Again, the factor would still be useful in representing students' complex problem solving abilities, but without making the (not yet supported) assumption that something like a general ability to solve complex problems exists.

Taken together, it becomes evident that research on CPS and its psychometric structure is still in its infancy and that substantial questions have to be investigated before clear conclusions about the construct's validity can be drawn. In systematically juxtaposing faceted and hierarchical measurement models when studying CPS's psychometric structure and relation to reasoning, we have made a bold and thorough first step in this direction. If it would be possible to adapt problem-solving scenarios to younger samples as well, the study of CPS's development could then shed further light on its validity. Let us hope that future research on CPS's construct validity will follow up on the issues raised here and bring additional insights. The free online publication of the Genetics Lab will hopefully also contribute to stimulating and enabling research on this topic. Although the discussion of whether CPS's facets and general ability correspond to real, existing, latent variables may seem like philosophical hairsplitting, a strong case for the existence of these constructs would be of paramount importance to back up the high usefulness of problem-solving scenarios in education (see above).

4.2.3. New chances and challenges for psychological assessment

In the course of the present dissertation, we tackled several issues that point to new chances but also to challenges of psychological assessment itself. Unfortunately, not all of these challenges could adequately be addressed within this project and can thus be seen as some of its limitations. First, while developing the Genetics Lab, it quickly became evident that

traditional test development procedures will not suffice. In 2.1., we argue that the development of such complex computer-based assessment instruments has to undergo extensive usability testing and suggest an extended test development process (see 2.1., Figure 3). Such a procedure not only assures acceptance of the instrument among the target population but also, and at least equally importantly, helps to avoid the idea that the test developer projects meaning into test responses that is simply not justified, a problem which is famously termed the "psychologist's fallacy" (cf. Markus & Borsboom, 2013, p.248).

Thorough usability testing ensures that the meaning of symbols, input modes, pictures, etc. is the same for the test developer as well as the test's target population. Unfortunately, despite the growing complexity of psychological tests, usability testing is only slowly being acknowledged in psychological assessment; for a notable exception, see for example Weinerth, Koenig, Brunner, and Martin (2014). However, including usability tests in the test development process is only a small part what could (and should) be considered in the future of psychological assessment. Although we included game-like characteristics in the GL in order to ensure high test motivation, we were still far beyond the possibilities of gamification (Dominguez et al., 2013) and in future, it may even be wise to consult game-developers when conceptualizing the assessment instrument.

We also repeatedly demonstrated the usefulness of process measures, based on so-called traces that are left by each student when interacting with the problem scenario. Study 4 clearly showed the potential of such process measures by revealing a more efficient problem exploration strategy of immigrant students; an advantage that was not found in all other administered measures. Added value was also shown, for example, in Study 2, in which we identified students that did not work properly on the GL by inspecting their time on task (for a comprehensive treatment of time on task and ability in complex problem-solving, see also Scherer, Greiff, & Hautamäki, 2015). In conventional tests, drawing on paper-pencil formats,

it is not possible to differentiate between a serious and a guessed response, at least not without sophisticated statistical treatment. Still yet, we are at the very beginning of understanding and fully exploiting the potential of process measures and there is strong need for further research (Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007; Winne, 2006).

Finally, we also gained experience concerning a timely form of test publication, since we published the Genetics Lab in three languages (English, French, German), including a user's guide and accompanying scientific documentation, freely online (www.assessment.lu/geneticslab). This procedure should not only guarantee full transparency concerning the applied measurement instrument, but should also inspire other researchers to conduct further studies on CPS, and allow educators to use the GL within their courses to gather experiences with CPS. In addition, we also aimed at increasing the project's scientific and public impact. As of April 2015, the time this thesis was completed, the GL had been downloaded more than a thousand times. With the exclusion of Antarctica, the GL was present on all continents and in countries ranging from Austria to Zimbabwe. Currently, there are ongoing studies (that we know of) in Austria, Australia, China, Italy, and the U.S., highlighting the impact the project has had on the field. The Genetics Lab was featured on psychologytoday.com, additionally increasing its popularity. In the meantime, a fourth language, Italian, was added to the published package.

This "success," however, comes at a cost and holds several risks. Maintaining the related website and responding to numerous (technical) queries is time-consuming and requires staff. Constantly adapting the test software due to the users' updated operating systems is a never-ending challenge in and of itself. Note that such continuous modifications are hardly necessary for paper-and-pencil-based tests. Luckily, the Luxembourg Centre for Educational Testing (LUCET, formerly EMACS) at the University of Luxembourg, under the direction of Prof. Dr. Romain Martin, provided a highly supportive infrastructure. But publishing a test also

entails responsibility for what is done with the test (cf. AERA, APA, & NCME, 2014). Although the GL is presented and published as a mere research tool, we do not know the precise purposes for which it is being used in each and every case. An adequate interpretation of the GL's performance scores should be facilitated by the accompanying scientific literature, but cannot be guaranteed. In addition, publishing the rationale of the GL's problem scenarios was necessary in terms of transparency and comprehensibility, but could endanger its validity, since the rational could become known among (future) test takers. Future projects thinking about using the effective channel of open online publication to increase transparency and visibility should be aware of these dangers.

References

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitzschke, A., & Funke, J. (2012). Dynamische Problemlösekompetenz. *Zeitschrift Für Erziehungswissenschaft*, *15*, 363–391.

Adey, P., Csapó, B., Demetriou, A., Hautamäki, J., & Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educational Research Review*, *2*, 75–97.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (Seventh). NJ: Prentice-Hall.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology*, *104*, 682–699.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review; Psychological Review*, *111*, 1061.

Burton, R., & Martin, R. (2008). L'orientation scolaire au Luxembourg: "Au-delà de l'égalité des chances… le gâchis d'un potentiel humain." In R. Martin, C. Dierendonck, C. Meyers, & M. Noesen (Eds.), *La place de l'école dans la société luxembourgeoise de demain* (pp. 165–186). Bruxelles, Belgium: DeBoeck.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptions for social exchange. In J. Barkow, L. Cosmides, & J. Tooby, *In: The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

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.

Dominguez, A., Saenz-de-Navarrete, J., de-Marcos, L., Fernandez-Sanz, L., Pagés, C., & Martinez-Herráiz, J. J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, *63*, 380–392.

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

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174.

Elliot Bennett, R., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing Complex Problem Solving Performances. *Assessment in Education: Principles, Policy & Practice*, *10*, 347–359.

Fischer, A., Greiff, S., & Funke, J. (2011). The Process of Solving Complex Problems. *The Journal of Problem Solving*, *4*, 19–42.

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

Funke, J. (2003). *Problemlösendes Denken* [Problem solving thinking]. Stuttgart, Germany: Kohlhammer.

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

Gottfredson, L. (1997). Mainstream science on intelligence (editorial). *Intelligence*, *24*, 13–23.

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., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74–83.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A New Assessment Perspective. *Applied Psychological Measurement*, *36*, 189–213.

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.

Güss, C. D., Tuason, M. T., & Gerhard, C. (2009). Cross-National Comparisons of Complex Problem-Solving Strategies in Two Microworlds. *Cognitive Science*, *34*, 489–520.

Gustafsson, J.-E. (2008). Schooling and intelligence: Effects of track of study on level and profile of cognitive abilities. *Extending Intelligence: Enhancement and New Constructs*, 37–59.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*, 107–124.

Hallinan, M. T. (1994). Tracking: From Theory to Practice. *Sociology of Education*, *67*, 79.

Heller, K. A. (1991). *Begabungsdiagnostik in der Schul- und Erziehungsberatung [Assessment of potential in school and educationa counseling]*. Bern, Switzerland: Verlag Hans Huber.

Hunt, E. (2011). *Human Intelligence*. New York: Cambridge University Press.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *10*, 631–639.

Klapproth, F., Glock, S., Böhmer, M., Krolak-Schwerdt, S., & Martin, R. (2012). School placement decisions in Luxembourg: Do teachers meet the Education Ministry's standards? *The Literacy Information and Computer Education Journal*, *1*, 765–771.

Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation [Intelligence assessment via computer simulations]*. Münster, Germany: Waxmann.

Kuhn, D. (2009). Do students need to be taught how to reason? *Educational Research Review*, *4*, 1–6.

Maaz, K., Baeriswyl, F., & Trautwein, U. (2013). Studie: “Herkunft zensiert?” Leistungsdiagnostik und soziale Ungleichheiten in der Schule [Study: “Origin censored?” Performance diagnostics and social inequality in school]. In D. Deißner (Ed.), *Chancen bilden* (pp. 185–305). Wiesbaden, Germany: Springer.

Markus, K., & Borsboom, D. (2013). *Frontiers of validity theory: Measurement, causation, and meaning*. New York: Routledge.

Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: Multination findings from the Programme for International Student Assessment (PISA). *Journal of Educational Psychology*, *104*, 1054–1073.

Martinez, M. E. (2000). *Education as the cultivation of intelligence*. Mahwah, NJ: Lawrence Erlbaum.

Martinez, M. E. (2013). *Future Bright: A Transforming Vision of Human Intelligence*. Oxford, UK: University Press.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-Solving Transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: Macmillan.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

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

OECD. (2014). *Pisa 2012 results: Creative problem solving: Students’ skills in tackling real-life problems (Volume V)*. Paris, France: OECD Publishing.

Preckel, F., & Baudson, T. G. (2013). *Hochbegabung - Erkennen, verstehen, fördern [Giftedness - recognizing, understanding, fostering]*. München, Germany: Beck.

Prensky, M. (2001). Digital Natives, Digital Immigrants Part 1. *On the Horizon*, *9*, 1–6.

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

Ridgway, J., & McCusker, S. (2003). Using Computers to Assess New Educational Goals. *Assessment in Education: Principles, Policy & Practice*, *10*, 309–328.

Schalke, D., Brunner, M., Geiser, C., Preckel, F., Keller, U., Spengler, M., & Martin, R. (2013). Stability and Change in Intelligence From Age 12 to Age 52: Results From the Luxembourg MAGRIP Study. *Developmental Psychology*, *49*, 1529–1543.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the Relation between Time on Task and Ability in Complex Problem Solving. *Intelligence*, *48*, 37–50.

Schleicher, A. (2006). Where immigrant students succeed: a comparative review of performance and engagement in PISA 2003. *Intercultural Education*, *17*, 507–516.

Shavinina, L. V. (2009). *International handbook on giftedness.* Amsterdam, The Netherlands: Springer.

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, *42*, 137–144.

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

Tapscott, D. (1998). *Growing up Digital: the Rise of the Net Generation*. New York: McGraw-Hill.

UNESCO. (2009). *Policy Guidelines on Inclusion in Education*. Paris, France: UNESCO.

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Weinerth, K., Koenig, V., Brunner, M., & Martin, R. (2014). Concept maps: A useful and usable tool for computer-based knowledge assessment? A literature review with a focus on usability. *Computers & Education*, *78*, 201–209.

Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist*, *41*, 5–17.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and Assessment in Schools* (Second). New York: Longman.

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

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, *29*, 18–29.

Acknowledgements

The project on which this dissertation is based has, to a large degree, been the effort of an excellent team at the University of Luxembourg, and I'm grateful to the many people who have helped make the Genetics Lab a reality. First and foremost, I would like to thank my primary supervisor, Prof. Dr. Martin Brunner, whose expertise in the field and methodological ingenuity have taught me a lot and have been a massive source of inspiration throughout the last years. I deeply enjoyed working together and regretted your leaving the UL, on a personal as well as professional level. I would also like to thank my second supervisor Prof. Dr. Romain Martin, who always unconditionally supported this project by fostering innovative and out-of-the box approaches to computer-based (!) psychological assessment and providing the perfect environment for this work. I'm very happy that this cooperation is continuing within several projects of the Luxembourg Centre for Educational Testing and hope that it will lead to several joint future projects.

The Genetics Lab and the present dissertation would not exist in its current form without the brilliance (and sweat) of Uli Keller and Ricky François. You're still keeping me breathless with your skills and I'm happy (and lucky) to have you as colleagues, thank you! Data collection for the first studies would not have been possible without the great technical support of Markus Scherer; I will always remember the basement of Bâtiment 11! Thank you!

I would also like to thank the cooperating partners of this project, above all Hélène Mayer, Cyril Hazotte, and Dr. Thibaud Latour from the former Centre de Recherche Henri Tudor, who translated my "psychologist's descriptions of a computer-based test" into a great and well-running program, now known as the Genetics Lab. The same goes for highly inspiring discussions and exchanges with my colleague Prof. Dr. Samuel Greiff and Prof. Dr. Joachim

Funke from the University of Heidelberg. Cooperating with you taught me a lot and I owe you all a big thank you!

The concept of the Genetics Lab, as well as my psychological health while working on this project significantly benefitted from great conversations with my (former and present) colleagues Kasia Gogol, Marius Wrulich, Daniela Schalke, Gina Wrobel, Danielle Hoffmann, Monique Reichert, Magda Chmiel, Stef Schäfers, Caroline Hornung, Tun Fischbach and lately, Enrico Micheli, André Kretzschmar, and Max Greisen. To all of you, a big thank you!

Last but not least, this dissertation would have never been possible without the cooperation, good will and support of the participating schools, teachers, and above all, the students who were forced into a laboratory to investigate fictive creatures. I thank all of you and have the modest wish that you all might benefit in some way from the Genetics Lab, as well as the related research.

A very special thank you goes to my great friend and former colleague, Carrie “On” Kovacs who has helped me and this project in countless ways. Forever SHL! ;-)