UNIVERSITY OF LUXEMBOURG
Educational Measurement and
Applied Cognitive Science (EMACS)

PUBLIC RESEARCH CENTRE HENRI TUDOR

Genetics Lab

A Computer-Based Microworld to Assess Complex Problem Solving

Theoretical Background &
Psychometric Evaluation

Philipp Sonnleitner
Ulrich Keller
Romain Martin
Cyril Hazotte
Hélène Mayer
Thibaud Latour
Martin Brunner

Fonds National de la
Recherche Luxembourg

UNIVERSITY OF LUXEMBOURG
Educational Measurement and
Applied Cognitive Science (EMACS)

PUBLIC RESEARCH CENTRE HENRI TUDOR

Luxembourg, April 2012

Preparation of this report was supported by the National Research Fund Luxembourg (FNR/C08/LM/06).

Please cite this report as follows:

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

IMPORTANT: The user agrees that any reports or published results obtained with the software will acknowledge its use by citing the following reference:

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

Questions concerning the content of this report, the application itself, etc. should be directed to:

Philipp Sonnleitner, MSc

University of Luxembourg
EMACS research unit

Campus Walferdange
L-7201 Walferdange
Luxembourg
Phone: (+352)466 644 9514
eMail: philipp.sonnleitner@uni.lu
URL: http://www.emacs.uni.lu/

Content

Preface

The Genetics Lab is a computer-based, psychometric sound microworld to assess complex problem solving behavior. This report gives basic information about the theoretical background and psychometric evaluation of the Genetics Lab. The Genetics Lab was developed within a cooperation between the University of Luxembourg and the Centre de Recherche public Henri Tudor. This cooperation was supported by the National Research Fund Luxembourg (FNR/C08/LM/06).

Please note that this report should not be considered as a complete test manual. Rather its purpose is to briefly summarize the theoretical grounds on which the Genetics Lab (GL) was developed. Moreover, this report presents first empirical evidence that the performance scores of the GL demonstrated good psychometric characteristics in terms of their reliability and external validity to intelligence measures and school grades. Note that some parts of this report have already been published in greater detail in Sonnleitner et al. (2012).

Together with the *Genetics Lab – User's guide to apply, configure and adapt the Genetics Lab* (included in the Genetics Lab download package), this report should ease the application of the Genetics Lab and thus foster its use in research and educational contexts.

A special thank goes to Ingo Schandeler for creating the many creatures inhabiting the Genetics Lab and to Eric Francois and Markus Scherer for their technical support and expertise.

April 2012, the authors

References:

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

Summary

General cognitive ability ("Intelligence") is among the most important and useful psychological constructs capable of predicting health and longevity, academic success, as well as success on the job. Nevertheless, a major shortcoming of typical intelligence measures is that these paper-pencil instruments use static problem formats with a relatively limited amount of complexity and that they generally do not provide information on test takers' problem representations. The present project aimed to overcome these severe limitations of typical intelligence tests by using computer-based complex problem solving scenarios as an alternative assessment instrument of students' general cognitive ability. To this end, we developed a new computer-based assessment instrument which is available in a German, French, and English translation: the Genetics Lab (GL).

To yield valid and reliable performance scores the construction rationale of the GL combined problem-solving research grounded in experimental psychology with well-established principles from individual differences research and psychometrics. More specifically, in the GL the task of the students is to examine how the genes of fictitious creatures influence their physical characteristics. The complexity of a creature (i.e., a problem) depends on the number of genes or characteristics, the number of connections between them, the kind of connection (positive or negative), and whether characteristics change without being affected by genes. The examination of each creature is split into two consecutive phases: the exploration phase and the control phase. In the exploration phase, students actively manipulate the creature's genes. The effects of their genetic manipulations on characteristics are displayed in diagrams. By carefully analyzing this information, students can draw conclusions about the connections and formulate hypotheses that can then be tested. Students document the knowledge they acquire about the relations between genes and characteristics in a database. The resulting causal diagram can be interpreted as the theoretical model developed by the student exploring the creature. In the final control phase, students have to manipulate the genes to alter the characteristics of organisms and reach specified target values. This phase requires the competencies of using a theoretical model to inform concrete actions and controlling the resulting outcomes. Students' behaviour while working on the GL is recorded in a detailed log-file. This information is used to validate whether students work properly on the GL and to derive three performance scores: (a) a process-oriented score reflecting how systematically students explored the creatures, (b) a score that measures students' level of acquired knowledge about the relations between genes and characteristics, and (c) another process-oriented score that reflects students' control performance to achieve specified target values.

The development of the GL was informed by a series of qualitative and quantitative small-scale studies where adults and adolescent students

participated. The results of these studies helped significantly to improve the conceptualization and lay-out of the computer-based assessment environment and the instructions that are given to test takers to allow for efficient working on the problems presented in the GL. These studies also showed that the GL enjoys a satisfactory level of acceptance among adolescent students and provided preliminary evidence that the performance scores of the GL demonstrated good psychometric characteristics in terms of their reliability and external validity to intelligence measures and school grades.

To scrutinize the psychometric characteristics of the GL we conducted a large-scale study where 300 ninth and 263 eleven graders participated who attended the intermediate and highest academic tracks of Luxembourgish schools. These are the grade levels at which many students make their transition from school to the workplace therefore making an assessment of students' general cognitive ability particularly relevant. The analyses of this rich data base are still on-going. First results showed that all performance scores of the GL demonstrated a satisfactory level of reliability. Further, results obtained from structural equation models indicated that complex problem solving may be either structurally conceptualized as a construct with three interrelated specific components (reflecting systematic exploration, system knowledge, and control performance) or as a hierarchical construct where (general) complex problem solving is at the apex and the more specific components of complex problem solving are located at the next lower level of the hierarchy. Moreover, complex problem solving and its various components were significantly related to key educational achievement indicators: (a) school grades in mathematics, science and languages, (b) test performance in mathematics, reading comprehension in German, reading comprehension in French as assessed by the national school monitoring in Luxembourg, and (c) performance on the PISA 2006 tests in mathematics, reading, and science. Notably, these data on students' educational attainment were collected at the same time when the GL was administered (i.e., grades), four months before (i.e., national school monitoring), or 2 years before (i.e., PISA 2006) suggesting that individual differences on the performance scores of the GL demonstrate a considerable level of temporal stability that is similar in magnitude to traditional measures of intelligence. Finally, complex problem solving (and its more specific components) could be clearly distinguished from reasoning ability (as measured by typical intelligence tests): although these constructs correlated substantively the correlations were far from perfect. Thus, both reasoning ability and complex problem solving represent important constructs to describe individual differences in cognitive performance among adolescent students. To study the incremental validity of GL performance scores over traditional reasoning measures we specified structural equation models that controlled for the common variance reflecting general cognitive ability. In doing so, our results showed that the impressive associations between complex problem solving and measures of educational attainment are largely attributable to a common variance core. Further, variance that is specific to complex problem solving measures may become particularly relevant for the

prediction of performance in technology rich environments. Finally, our results showed that the GL is measurement invariant across student groups with varying migration background suggesting that the GL provides a fair assessment of students' level of CPS.

To conclude, the results of our empirical studies showed that the GL enjoys a satisfactory level of acceptance among adolescent students and that it may be considered as reliable, valid, and fair alternative assessment instrument of students' general cognitive ability by means of CPS scenarios. Future projects may therefore use the GL for the assessment of individual students' problem solving strategies and knowledge representations as well as for the training of students' problem-solving abilities.

Part I - Theoretical framework

1.1. Why do we need an alternative measure of GCA?

The GL has been developed to assess students' GCA (i.e., "intelligence") because conventional intelligence tests faced criticisms that they tend to neglect that many problems are dynamic, entail problem recognition, require to look for necessary information in order to acquire new knowledge, and require motivation and personal involvement (cf. Sternberg, 1985). Hence, in light of these limitations of conventional intelligence tests, alternative measures of GCA are needed. To this end, we developed a new computer-based assessment environment—a so-called microworld— to measure GCA: the Genetics Lab (GL). In doing so, we also aimed to provide a reliable and valid instrument that responds to the requirements of many contemporary educational curricula and educational assessment frameworks (OECD, 2004, 2010) that emphasize the critical importance of the (domain-general) ability to solve complex problems (e.g., Ridgway & McCusker, 2003).

1.2. Conceptualization 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

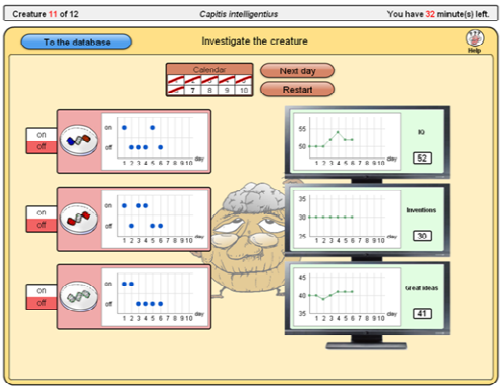
(number and type of the underlying connections), the degree of their "eigendynamic" (change of variables without intervention; see Frensch & Funke, 1995), opaqueness (the underlying connections are hidden) and multiple goals (number of output variables which must be influenced). In order to distinguish between knowledge acquisition skills and knowledge application skills, working with such a scenario is divided into an "exploration" phase and a "control" phase. Further, the GL also capitalizes on a current methodological advancement within the DYNAMIS tradition—the MicroDYN-approach (Greiff, Wustenberg, & Funke, 2012)—that combines problem-solving research grounded in experimental psychology (as described above) 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.

1.3. Task requirements 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).

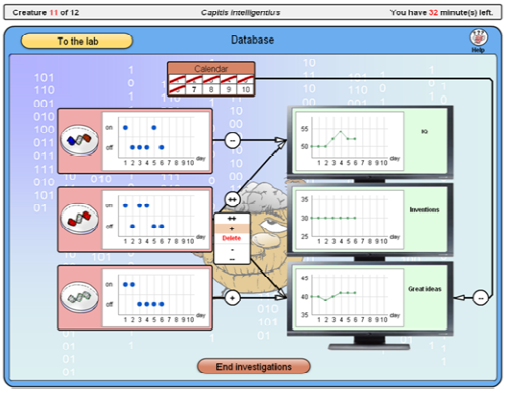
As noted above, the complexity of a creature depends on (a) the number of genes or characteristics, (b) the number of connections between them, (c) the kind of connection (positive or negative), and (d) whether characteristics change without being affected by genes (eigendynamic). The examination of 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) and document their knowledge in a database (Figure 1b). In the control phase, students are required to manipulate the genes to achieve specified target values on certain characteristics (Figure 1c). Students' behavior while working on the GL is recorded in a detailed log-file which may be saved on a local storage medium or on an external server via the internet. The log-file is then used to derive performance scores as well as to validate whether students worked properly on the GL. Specifically, the log-file allowed us to develop a new process-oriented score (*Systematic Exploration*) reflecting how systematically students explored the creatures. Exploration is most informative for solving the task if students manipulate the genes in a way that changes in characteristics can be unambiguously attributed to a certain gene (see Kröner, Plass, & Leutner, 2005; Vollmeyer, Burns, & Holyoak, 1996). At any time during the exploration phase, students can document their knowledge in a database (Figure 1b). We scored these records by modifying an established scoring algorithm (see Funke, 1992). The resulting *System Knowledge* score reflects

a. Phase 1: Exploring the Creature



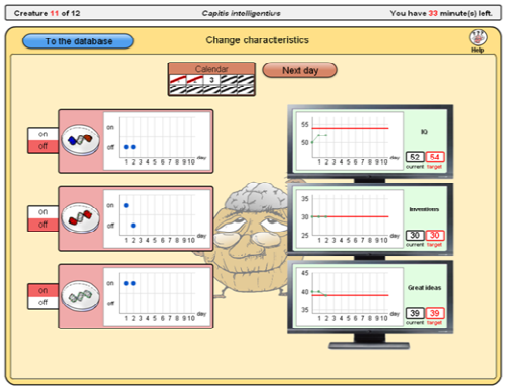
Students explore the effects of genes on certain characteristics of a number of creatures in a fictitious lab. By manipulating genes and observing the characteristics for a certain time, students can draw conclusions about the connections and formulate hypotheses that can then be tested.

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 creatures 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.

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).

knowledge about how a gene affects a certain characteristic of a creature and knowledge about the strength of such an effect.

During the control phase students can look up the correct relations between genes and characteristics which are provided in the data base. Note that the required 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 developed a new scoring algorithm taking into account the quality of each applied control step. Every step is scored whether it is optimal or not with regard to the achievement of the target values. A step is optimal if it maximally decreases the distance to the target values during control phase. As there are three steps per control phase of each item, a maximum of 3 is possible per item.

1.4. 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 newly developed performance scores. First, many previous microworlds were based on a single but very extensive problem scenario. This so-called one-item approach has severe shortcomings (Greiff et al., 2012; Kröner, Plass, & Leutner, 2005): (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. The GL, in contrast, is based on the MicroDYN approach (Greiff et al., 2012), in which students examine several independent scenarios (i.e., several creatures). 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 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. Rollett, 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, 2003; 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, 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. Finally, most previous microworlds are available in a single language only. The GL, on the other hand, was translated into three different languages (i.e., English, German, and French); students can freely choose the language in which they want to work on the GL. Doing so, should guarantee fullest possible understanding of instructions and test requirements among student test takers in the multicultural and multilingual educational environment of Luxembourg.

Part II - Acceptance of the Genetics Lab among adolescent test takers

In applied assessment, it is essential that the instruments administered are accepted by the test takers (and by those who use the scores obtained). 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 u. a., 2005) or even more decades (Omodei & Wearing, 1995; Vollmeyer et al., 1996). Moreover, to the best of our knowledge, the acceptance of these microworlds by student test takers has not yet been empirically investigated. To fill this research gap, we conducted an empirical study that examined students' acceptance of the GL.

2.1. Method

2.1.1. Participants and procedure

Participants were 61 ninth graders of an intermediate-track secondary school in Luxembourg. 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. 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).

2.1.2. 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 in the paper by Sonnleitner et al. (2012). Item scores were summarized to total scores where 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 considered 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.

Complex Problem Solving. The GL was administered without a time limit and contained 16 scenarios of varying complexity. Performance scores were computed as described above.

2.2. Results and discussion

As shown in Table 1, Students rated the GL and its elements to be attractive and working with it to be fairly easy. 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 and functionality 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. 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 of subsequent versions of the GL that we also used in the second pilot study and in the main study (see below). In sum, these results provided initial empirical evidence that the GL can be applied in an educational context, where student acceptance is considered to be important.

Table 1. Acceptance of the GL and relation of acceptance with performance scores

Acceptance Score							Correlation with CPS scores		
(Number of Items)	α	M	SD	p25	MD	p75	SE	SK	CP
Perceived Ease of Use (4)	.71	54	23	44	56	69	.31	**.44**	**.39**
Attractivity (9)	.91	64	22	56	67	78	.22	**.34**	**.54**
Comprehensibility (10)	.81	61	17	50	63	73	.28	**.49**	.29
Functionality (7)	.82	60	22	46	64	71	.17	**.35**	**.56**

Note. All correlations printed in bold are significant at $p < .05$ (2-sided testing). α = Cronbach's alpha.; p25 = first quartile (Q1); p75 = third quartile (Q3). SE = Systematic Exploration; SK = System Knowledge; CP = Control Performance

Part III - Psychometric characteristics of the Genetics Lab

Is the GL a reliable and valid measure of students' ability to solve complex problems? To answer this question, we drew on a large student sample to investigate the internal consistency of the performance scores of the GL as a measure of their reliability. Further, to study the construct validity of these scores we analyzed their factorial structure to clarify vital questions on the nature of CPS: Is CPS best represented as a single, a multidimensional or a hierarchical construct? We also analyzed the convergent validity of the GL with respect to

traditional measures of GCA, namely paper-pencil tests of reasoning ability. Further validity evidence was obtained from analyzing the relations of the GL scores to key indicators of academic success, namely school grades, and performance on domain-specific competency tests as applied in the Luxembourgish school monitoring program and in the Programme for International Student Assessment (PISA). Finally, one possible application of the GL is the national school monitoring program in Luxembourg where a large portion of the students have a migration background. Thus, we studied whether the GL is a fair assessment of students' level of CPS irrespective of their migration background.

3.1. Method

3.1.1. Participants and procedure

Participants in the main study were 563 students who were enrolled in schools in Luxembourg: 300 students attended the 9th grade (146 females; age: M = 15.6 years; 112 of these ninth-graders attended the highest academic track). The remaining 263 students attended the 11th grade (138 females; age: M = 17.4 years; 217 were enrolled in the highest academic track). Moreover, we adopted the definition of migration background that is used in PISA. Doing so, indicated that 375 students had no and 185 students had some migration background (for 3 students this information was not available). All study materials (i.e., the GL, intelligence tests, and a questionnaire) were administered by trained test administrators within the time constraint of two school hours (i.e., about 100 minutes)

3.1.2. Measures

***Complex Problem Solving*.** Students were given standardized on-line instructions how to use the GL and worked on practice scenarios. This instruction period lasted 15 minutes. Afterwards, students completed 12 GL scenarios of varying complexity within an overall time limit of 35 minutes which allowed the vast majority of students to complete all scenarios. For many analyses we summarized the performance across scenarios by computing three total scores (as described above): (a) Systematic Exploration (SE), (b) System Knowledge (SK), and (c) Control Performance (CP). Moreover, for some analyses, we summarized performance of a smaller number of scenarios by means of parcel scores (E1 to C3; Figure 2). To this end, we applied the shared-uniqueness strategy advocated by Hall, Snell, & Singer Foust (1999).

Reasoning Ability. Reasoning ability was measured by three subtest scores that assessed students' ability (a) to complete figural matrix patterns (time limit: 10 minutes; score MA, Figure 2), (b) to solve number series (10 minutes; score NS), and (c) to mentally arrange geometric figures (7 minutes; score GF).

These subtests were taken from the Intelligence Structure Test IST-2000R (Amthauer, Brocke, Liepmann, & Beauducel, 2001), a reliable and valid measure of intelligence.

School Grades. Students reported subject-specific grades that they received in their last report card in (a) mathematics, (b) science, (c) French, and (d) German. These grades may range from 0 to 60 with higher grades indicating better achievement.

Achievement Tests. A large proportion of the students in grade 9 (n = 285) also participated in the year 2011 cycle of ÉpStan, the national school monitoring program. Hence, for these students data were available on computer-based competency tests for German reading comprehension, French reading comprehension, and mathematics. Moreover, 87 students in grade 11 participated in the year 2009 cycle of PISA; for these students data were available on paper-pencil tests for reading, mathematics, and science.

3.1.3. Statistical analyses

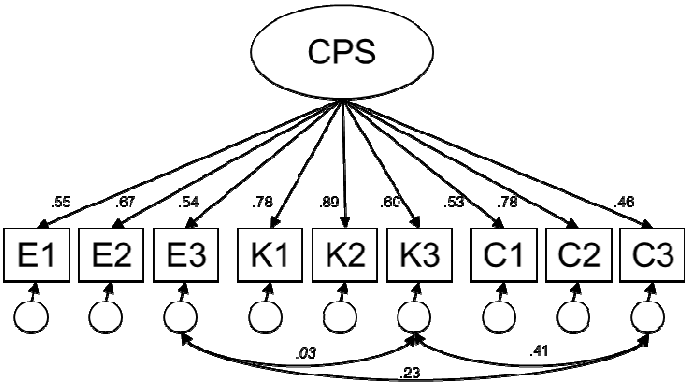
We embedded the statistical analyses in the structural equation modeling environment (a) to capitalize on recent psychometric advances in confirmatory factor analysis (e.g., Eid, Lischetzke, Nussbeck, & Trierweiler, 2003), (b) to efficiently handle the observed missing data patterns on reasoning measures and on achievement tests, and (c) to use multiple-group models to scrutinize whether the applied cognitive measures are invariant across student groups with differing migration background. Model parameters were estimated with Mplus 5.2 (Muthén & Muthén, 1998-2007). To evaluate model fit we consulted several descriptive measures of model fit that are recommended in the literature: the Standardized Root Mean Squared Residual (SRMR), the Comparative Fit Index (CFI), and gamma, which is based on the more popular Root Mean Square Error of Approximation (RMSEA). Compared to the RMSEA, gamma has the advantage that it provides a more realistic test of model fit when the number of manifest variables is small (Fan & Sivo, 2007): SRMR values below .08, CFI values above .95, and gamma values above .95 are generally considered to indicate good model fit (Hu & Bentler, 1999).

3.2. Results and discussion

3.2.1. Structure of complex problem solving

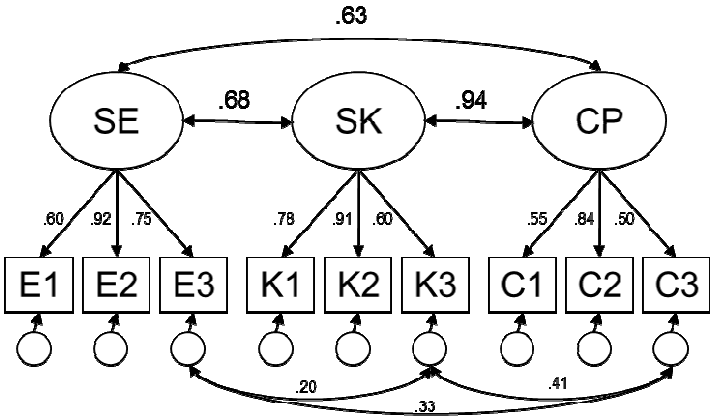
Is CPS best represented as a single, a multidimensional or a hierarchical construct? A general factor model that represented CPS as a single construct was not supported by our data (Figure 2a). Rather the evaluation of model fit

a



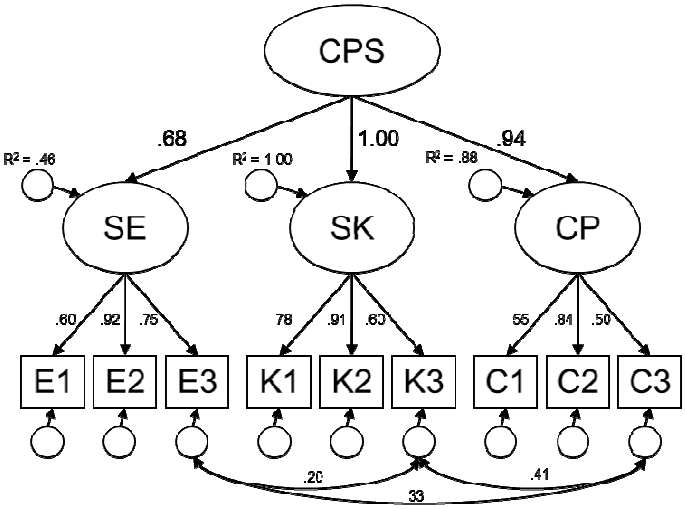
$\chi^2(24, N = 563) = 404$, $p < .001$
CFI = .84, SRMR = .07, RMSEA = .17, gamma = .87

b



$\chi^2(21, N = 563) = 117$, $p < .001$
CFI = .96, SRMR = .04, RMSEA = .09, gamma = .96

c



$\chi^2(22, N = 563) = 118$, $p < .001$
CFI = .96, SRMR = .04, RMSEA = .09, gamma = .96

Figure 2. Psychometric evaluation of the Genetics Lab (GL): Structure of Complex Problem Solving in terms of (a) single construct, (b) a multidimensional construct or (c) a hierarchical construct. Standardized model parameters are shown in Figures a, b, and c. CPS = complex problem solving; SK = system knowledge; SE = systematic exploration; CP = control performance.

showed that CPS may be conceived as a multidimensional (Figure 2b) or a hierarchical construct (Figure 2c). Notably, SE, SK, and CP were found to be substantially intercorrelated (Figure 2b). Although the correlation between SK and CP was very high, further analyses showed that this correlation was statistically different from $r = 1$. Thus, these results support the conclusion that individual differences in SK and CP represent distinct components of students' CPS behavior. Moreover, we found that those parcel scores that were based on scenarios where students had to detect eigendynamics shared variance over and above the latent target constructs. The results of the higher-order factor model supported the idea that the substantial correlations between the more specific components of CPS are indicative of the operation of a higher-order construct that may be interpreted as students' general ability to deal with complex problems. Interestingly, system knowledge was found to be perfectly related to the general CPS construct. These results suggest that building up system knowledge is central to CPS: it represents both the outcome of students' systematic exploration processes as well as the building block for successful control performance. In sum, both the multidimensional and hierarchical conceptualization of CPS may be useful in future research. The choice of these conceptualizations should be guided by the question whether the specific components or the global aspect of CPS is the major focus of research.

3.2.2. Reliability and validity of the Genetics Lab

Crucially, reliability coefficients (in terms of Cronbach's Alpha) for the various performance scores ranged between .79 (CP) and .91 (see top panel of Table 2). Hence, our reliability analyses showed that the performance scores of the GL (which are currently based on 12 scenarios) can be considered to be reliable measures of the specific components of CPS and general CPS itself. When higher score reliabilities are needed (e.g., for high-stakes selection decisions) one can easily add more scenarios to the GL by taking advantage of the construction rationale of the GL that allows to easily generate more scenarios with fairly well predictable levels of difficulty (e.g., by manipulating the number of genes).

One important line of evidence to support the construct validity of the GL as a measure of GCA is its relation to reasoning ability. Reasoning ability lies at the heart of many conceptualizations of GCA and reasoning tests are also included in almost all intelligence test batteries which renders the empirical study of the association between scores of the GL and these traditional measures of GCA particularly important. Importantly, the results obtained for reasoning ability (see Table 2) may also serve as an empirical benchmark to evaluate the psychometric properties of the GL. We used a latent variable model (not shown in Figure 2) for these analyses to correct for the disattenuation of correlations due to measurement error of the applied measures. In line with several previous studies (e.g., Danner, Hagemann, Schankin, Hager, & Funke, 2011; Kröner et al., 2005;

Wüstenberg, Greiff, & Funke, 2011), our results showed that reasoning ability is significantly related with general CPS and its specific components (Table 2). Yet, these correlations are clearly different from $r = 1$ suggesting that reasoning ability and CPS may tap different aspects of GCA. To gain further insights into the validity of the GL we also examined the relations of students' CPS performance to key indicators of academic success. To this end, general school achievement was measured as a common factor reflected by school grades, and the domain-specific measures applied in ÉpStan or PISA, respectively. Our results showed that general CPS and the specific components of CPS are significantly related to all indicators of academic success.

Table 2. Reliabilities and relations of the performance scores of the GL to reasoning ability and achievement measures

	Reasoning	SE	SK	CP	CPS
		Reliability			
Cronbach's Alpha	**.88**	**.91**	**.90**	**.79**	**.89**
		Correlations with External Criteria			
Reasoning Ability	---	**.38**	**.56**	**.59**	**.62**
Mathematics Achievement					
Mathematics Grade	**.33**	**.16** (.04)	**.22** (.04)	**.30** (.11)	**.25** (.05)
ÉpStan	**.64**	**.30** (.08)	**.52** (**.17**)	**.59** (**.24**)	**.58** (**.22**)
PISA	**.52**	**.30** (.11)	**.31** (.01)	**.40** (.08)	**.40** (.06)
French Achievement					
French Grade	**.12**	**.09** (.06)	.07 (.02)	.03 (-.05)	.05 (-.03)
ÉpStan	**.16**	**.32** (**.27**)	**.35** (**.29**)	**.37** (**.32**)	**.39** (**.37**)
German Achievement					
German Grade	**.27**	**.12** (.02)	**.26** (**.11**)	**.25** (.09)	**.27** (**.12**)
ÉpStan	**.48**	**.22** (.02)	**.50** (**.23**)	**.52** (**.24**)	**.56** (**.31**)
PISA	**.37**	**.23** (.08)	**.30** (.09)	**.25** (.02)	**.28** (.05)
Science Achievement					
Science Grade	**.42**	**.17** (.01)	**.30** (.05)	**.33** (.08)	**.31** (.06)
PISA	**.54**	**.31** (.11)	**.39** (.10)	**.41** (.08)	**.43** (.11)
General School achievement					
Grade Point Average	**.50**	**.22** (.03)	**.38** (.09)	**.40** (.10)	**.39** (.10)
ÉpStan	**.74**	**.38** (.11)	**.68** (**.28**)	**.75** (**.35**)	**.76** (**.38**)
PISA	**.55**	**.31** (.10)	**.40** (.09)	**.42** (.09)	**.43** (.13)

Note. Correlations are corrected for measurement error; numbers in parentheses are partial correlations of performance scores to external criterial controlling for the common variance with reasoning ability. Numbers in bold print are statistically different from zero (at α =. 05; two-sided testing).

Notably, performance on the achievement tests of ÉpStan and PISA was assessed 4 months and almost two years before the GL was administered, respectively. Hence, these results empirically underscore that individual differences on performance scores of the GL possess postdictive validity as well as a considerable level of temporal stability. Finally, to learn about the incremental ability of CPS scores (i.e., do the scores of the GL predict academic success over and above reasoning ability?) we specified a confirmatory factor model that controlled for the variance of the scores of the GL that is shared with reasoning ability (see Eid et al., 2003; the model is not shown in Figure 2). The key results of these analyses are displayed in parentheses in Table 2. Controlling for reasoning ability resulted in a considerable drop in correlation coefficients between GL scores and indicators of academic success. Importantly, the pattern of results also pointed to the operation of a method effect attributable to the mode of test administration. When the indicators of academic success were paper-pencil based (which concerns the PISA tests and school grades which are largely based on written examinations) the correlations between performance scores of the GL and academic success approached zero. However, when the indicators of academic success were also computer-based (i.e., ÉpStan test scores) performance scores of the GL demonstrated incremental validity. Taken together, the present pattern of results strongly supported the construct validity of the GL as a measure of GCA. Scores of the GL were significantly related to reasoning ability–the traditional measure of GCA. Likewise to reasoning tests, scores of the GL were found to be significantly related to key indicators of academic success. Moreover, relations of the scores of the GL to academic success reflected to a large degree associations of these indicators with reasoning ability. Hence, it is the common variance of reasoning ability and GCA that largely accounts for the observed relations of the scores of the GL to indicators of academic success. Moreover, performance scores of the GL demonstrated incremental validity (over and above paper-pencil measures of reasoning ability) for computer-based measures of academic success which suggests that scores of the GL may be a useful alternative measure of GCA when it comes to the prediction of performance in technology rich environments.

3.2.3. Test fairness of the Genetics Lab

Finally, one possible application of the GL is the school monitoring program in Luxembourg where a large portion of the students have a migration background. We therefore studied whether the measurement properties of the GL and the reasoning tests are invariant across students with differing migration background. To this end, we used a stepwise approach based on multiple-group factor-analytic models (Little, 1997; Lubke, Dolan, Kelderman, & Mellenbergh, 2003). Our results (see Table 3 and Figure 3) indicated that a strict invariant model specification provided a good fit to the data which strongly supported the assumption that the scores of the GL (as well as the reasoning measures) are indeed measurement invariant. Moreover, students without migration background

outscored students with migration background both on reasoning and on general CPS. Further analyses by means of a MIMIC model showed that these mean differences became negligibly small when taking into account that students with migration background were more likely to be enrolled in lower academic tracks and lower grade levels (see negative correlations in Figure 3 MIMIC model). Interestingly, grade level and academic track explained a considerably larger portion of variance in reasoning ability than in general CPS. In sum, these results provided strong support to consider the GL as a fair assessment of students' level of CPS irrespective of their migration background. Moreover, given that the GL is less affected by differences in educational background than reasoning tests, the GL might be a better measure of students' level of cognitive potential than typically applied paper-pencil reasoning tests.

Table 3. Fit statistics of different measurement models testing measurement invariance of the Genetics Lab

Model	χ^2	df	p	CFI	SRMR	RMSEA	gamma
Models to Test Measurement Invariance							
Configural	34	16	.01	.98	.03	.06	.98
Metric	38	20	.01	.98	.05	.06	.98
Scalar	41	24	.02	.98	.05	.05	.98
Strict	47	30	.02	.98	.05	.05	.98
Adjustment of Mean Differences Due to Migration Background for Track and Grade Level							
MIMIC	32	20	.04	.99	.02	.03	1.00

Note. χ^2 = chi-square goodness-of-fit statistic; df = degrees of freedom; CFI = comparative fit index; SRMR = standardized root-mean square residual; RMSEA = root mean square error of approximation

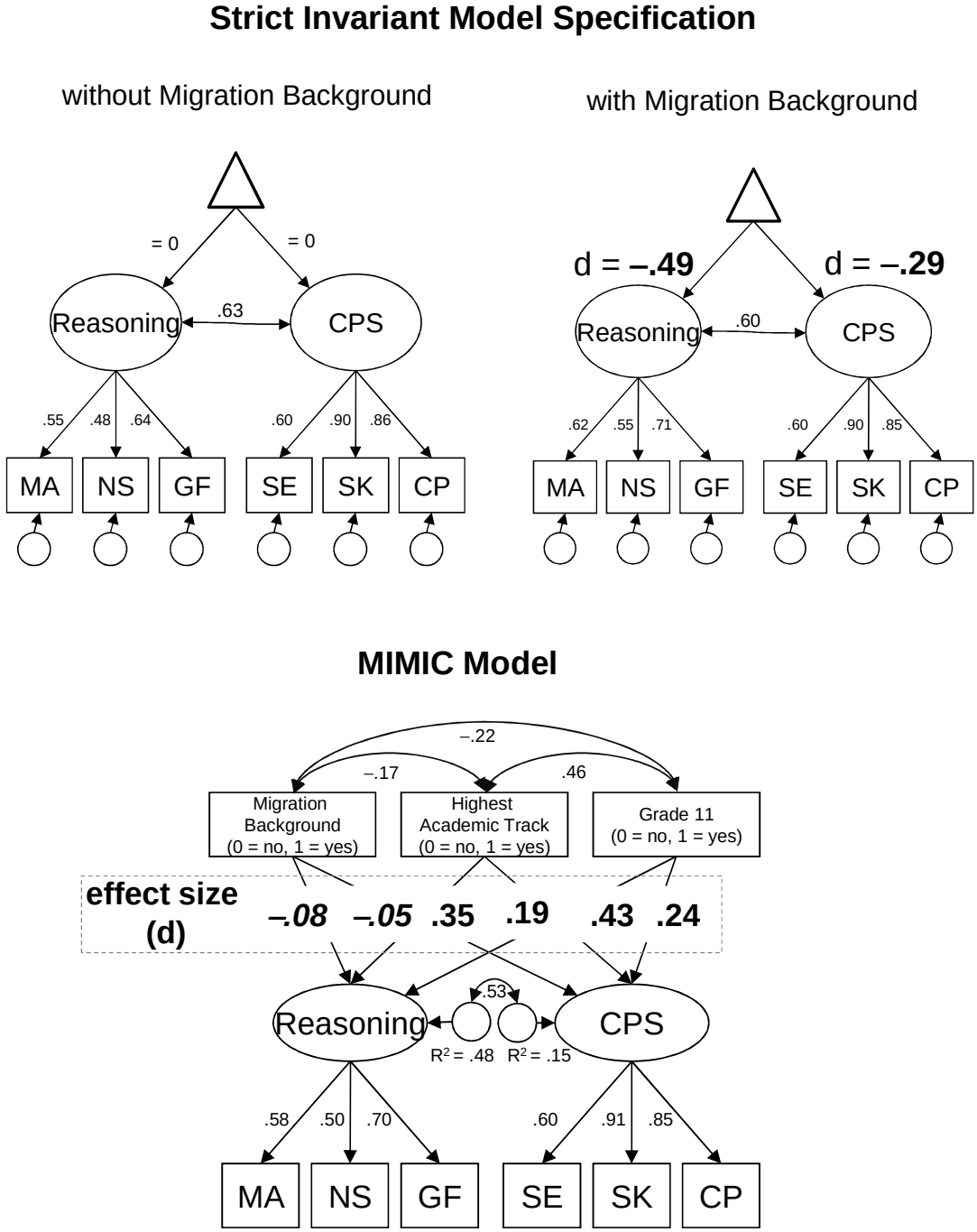


Figure 3. Testfairness of the GL with respect to migration background: Measurement invariance and adjustment of mean differences due to migration background for differential enrollment in academic tracks and grade levels (by means of a MIMIC model). Standardized model parameters are given. CPS = complex problem solving; SK = system knowledge; SE = systematic exploration; CP = control performance. MIMIC = Multiple Indicators Multiple Causes

References

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

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.

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 CT-C (M-1) model. *Psychological Methods*, *8*, 38.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*, 509–529. doi:10.1080/00273170701382864

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

Funke, J. (1993). Microworlds based on linear equation systems: A new approach to complex problem solving and experimental results. Elsevier Science Publishers.

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

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

Frensch, P. A., & Funke, J. (1995). Definitions, traditions and a general framework for understanding complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving. The European perspective* (pp. 3-26). Hillsdale: Erlbaum.

Greiff, S., Wustenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new assessment perspective. *Applied Psychological Measurement*, *36*, 189–213. doi:10.1177/0146621612439620

Hall, R. J., Snell, A. F., & Singer Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, *2*, 233–256. doi:10.1177/109442819923002

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

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

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

Little, T. D. (1997). Mean and Covariance Structures (MACS) Analyses of cross-cultural data: practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*, 543–566.

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and instruction*, *13*, 125–139.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(43-52.

McPherson, J., & Burns, N. R. (2007). Gs Invaders: Assessing a computer game-like test of processing speed. *Behavior research methods*, *39*, 876–883.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide (5th ed)*. Los Angeles: CA: Muthén & Muthén.

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

OECD (2010). *PISA 2012 Field Trial Problem Solving Framework. Draft Subject To Possible Revision After The Field Trial.* Paris: OECD (Retrieved June 15, 2011, 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.

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

Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education: Principles, Policy & Practice*, *10*, 309–328. doi:10.1080/0969594032000148163

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: Lit Verlag.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., u. a. (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.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

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.

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

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

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods*, *35*, 185–193.

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*, *7*, 1–10.

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