

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

Philipp Sonnleitner¹, Martin Brunner², Samuel Greiff³, Joachim Funke³, Ulrich Keller², Romain Martin², Cyril Hazotte⁴, H el ene Mayer⁴ & Thibaud Latour⁴

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 attractiveness) 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,

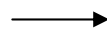
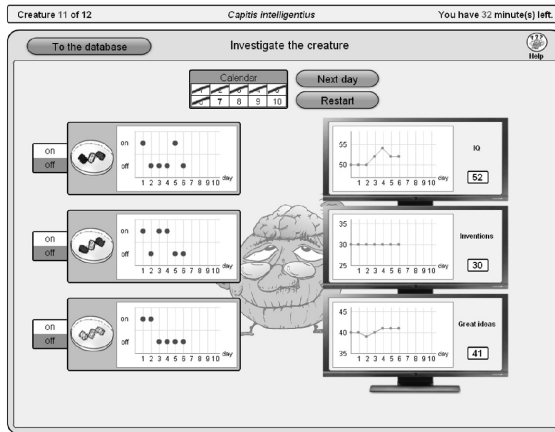
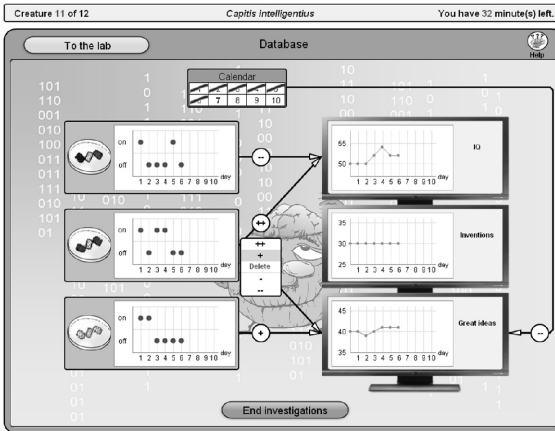


Figure 1: Screenshots of the different phases of the Genetics Lab: (a) Students explore how genes affect the characteristics of a fictitious creature and (b) record their knowledge in a database. (c) Students aim at achieving a given level of a characteristic (indicated by a red line and target value).



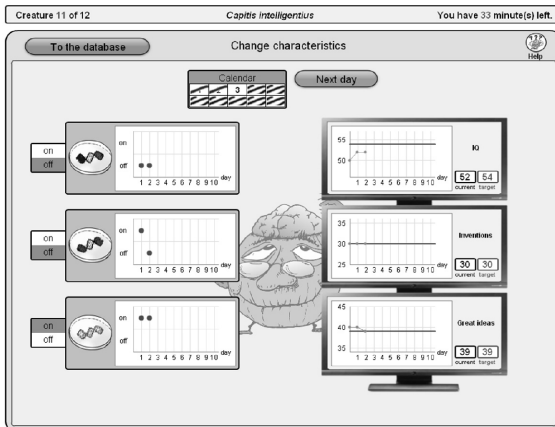
a. Phase 1: Exploring the creature

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 *Systematic 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 (*Control 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 $\alpha = .80$ (*Control Performance*) and $\alpha = .94$ (*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 $\alpha = .61$ (*Control Performance*) to $\alpha = .88$ (*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 Research*, 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. <http://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](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. http://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](http://dx.doi.org/10.1016/S0160-2896(02)00121-6)
- Rollett, W. (2008). *Strategieinsatz, 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 minimum (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