Running Head: ASSESSING CPS IN THE CLASSROOM

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

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

Abstract

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

Introduction

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

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

This chapter identifies and discusses challenges that arise when microworlds are administered “in the classroom”: the special characteristics of today’s students, also described as “digital natives”, and the need of timely behaviour-based scoring procedures that are at the same time easy to understand by educators and teachers. By taking the Genetics Lab, a microworld especially targeted at students at age 15 and above of all academic tracks as an example (Sonnleitner et al., 2012a), opportunities to react on these challenges are presented and evaluated on the basis of three independent studies using the Genetics Lab.

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

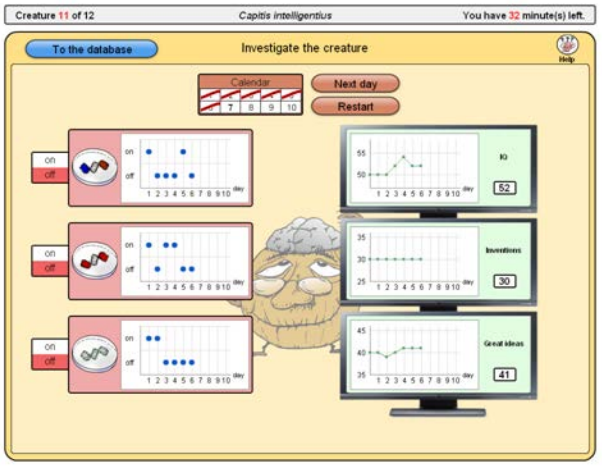
To learn more about the application of microworlds in the educational setting and to further investigate to what extent and in what way microworlds have to be adapted for this context, the Genetics Lab (GL) was developed at the University of Luxembourg (see Sonnleitner et al., 2012a, 2012b). The goal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

In the GL (shown in Fig. 1), students explore how genes of fictitious creatures influence their characteristics. To this end, students can actively manipulate the creatures’ genes by switching them “on” or “off” and then study the effects of these manipulations on certain

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

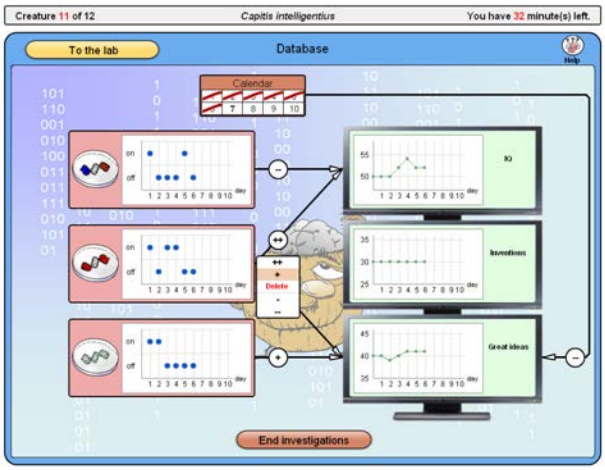
characteristics of the creatures (Fig. 1, a). Genes (i.e. input variables) are linked to the characteristics (i.e. output variables) by linear equations. It is the task of the student to find out about these (non-transparent) relations and to document the gathered knowledge (Fig. 1, b). Finally, the students have to apply the gathered knowledge to achieve certain target values on the creatures' characteristics (Fig. 1, c). These task characteristics allow to derive performance scores about (a) the students' exploration and information gathering behaviour, (b) the students' gathered knowledge in the form of a causal diagram showing the discovered relations between genes and characteristics, and (c) the students' ability to apply the knowledge in order to achieve certain target values on the creatures' characteristics. Each creature is designed in a way to realize key features of a complex problem (see e.g. Funke, 2001; Funke, 2003): (a) *complexity*, by including a high number of variables (several genes and characteristics), (b) *connectivity*, by linking the variables via linear equations, (c) *dynamics*, by implementing an automatic change of certain characteristics that is independent from the students' actions, (d) *intransparency*, by hiding the connections between the variables, and (e) *multiple goals*, by asking the student to achieve different target values on several of the creature's characteristics. For further details about the GL's scores and construction rationale please see (Sonnleitner et al., 2012a).

The GL has been applied in three independent studies so far with more than 600 participating students (see Table 1 for an overview). To foster commitment and motivation, detailed written feedback on the performance was offered. Further details concerning Study 1 and 2 are given in (Sonnleitner, et al., 2012a), concerning Study 3 in (Sonnleitner et al., 2012b). The gathered data along with the experiences made within these studies inform the following discussion of challenges and opportunities of microworlds within the educational field.



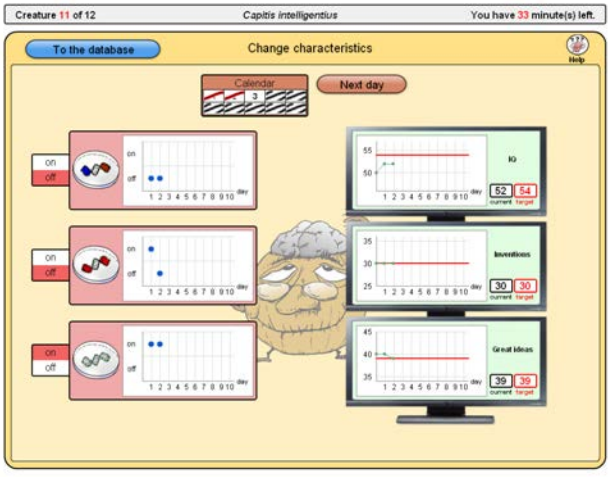
a.: Gathering Knowledge

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



b.: Documenting Knowledge

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



c.: Applying Knowledge

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

Figure 1: Screenshots of the different phases students have to go through within the Genetics Lab (taken from Sonnleitner et al., 2012a)

Table 1: Sample characteristics of the presented studies

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

Challenge 1 – Digital natives

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

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

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

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

This, in turn, has several crucial implications concerning the expectations of today's students with regard to a computer-based test: First, due to their massive exposure to high quality (commercial) computer programs, they expect a perfect and flawlessly functioning technology. Second, especially on the basis of the experiences made with video games and newer mobile devices, a completely intuitive graphical user interface (GUI) is expected. Students do not want to invest time and effort to figure out how to interact with a program. Third, this GUI should also be appealing and resemble modern standards of design to ensure that the test is perceived as being attractive and of high quality. Fourth, students want to learn how to deal with the task by

actively exploring and interacting with it. In contrast, extensive written instructions are very likely skipped by them. Finally, motivation to interact with a test will be high when they get instant gratification or at least instant feedback on their performance. If these criteria are not met by a computer-based test, acceptance of the instrument might be at stake.

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

A first step to respond to these special characteristics of today's students concerned the theoretical conceptualization of the GL. To start with, we ensured a clear and intuitive GUI by following recommendations of user interface design (e.g. Fulcher, 2003). As can be seen in Fig. 1, the structure of the GUI clearly resembles the navigation within the GL: The layer at the top shows the progress within the test itself by indicating the number of creatures (i.e. items) that are left and the remaining time for investigating them. The next layer corresponds to navigating within each item; it provides the buttons to switch between the lab and the database and contains the help-function. Finally, all elements to directly manipulate the creature or depict the gathered knowledge about it are arranged within the inner layer of the GUI. In addition, elements belonging together (e.g. the calendar and the buttons to progress in time) share the same colour. To make the design of the GL even more appealing and to increase the motivation to work on the test, we implemented several game-like characteristics (see McPherson & Burns, 2007; Washburn, 2003; Wood, Griffiths, Chappell, & Davies, 2004): A "cover story" was created, putting the student into the role of a young scientist that starts working in a fictive genetics lab. An older scientist charges the student with the mission of investigating several newly discovered creatures and explains the functioning of the lab. Throughout the test, this "virtual mentor" remains present in the form of an integrated help function. In addition, the fictitious creatures are depicted in a funny cartoon-like style and carry humorous characteristics (Fig. 1 and 2). Hence,

after exploring and manipulating the creature, the student gets performance contingent feedback in the form of two simple scales scoring the depicted causal diagram of phase 1 and the student's control performance of phase 2 (Fig. 2).

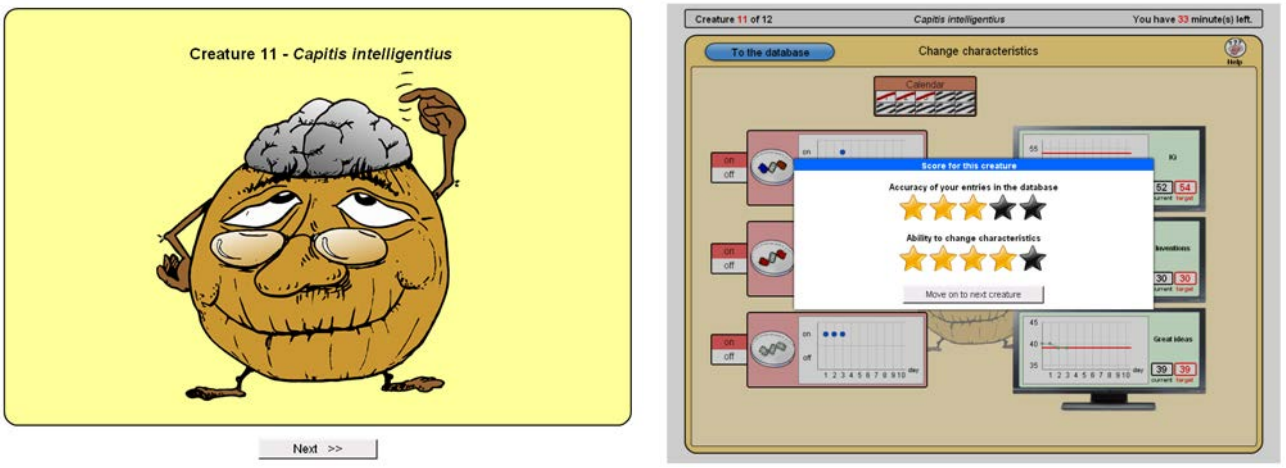


Figure 2: Start screen of creature (i.e. item) 11 (left side) and performance contingent feedback at the end of an item (right side).

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

The second step to ensure acceptance of the GL among digital natives was to guarantee a flawless functioning test of high usability. To this end, we adapted and substantially extended traditional test development procedures by including several small scale usability studies (Fig 3).

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

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

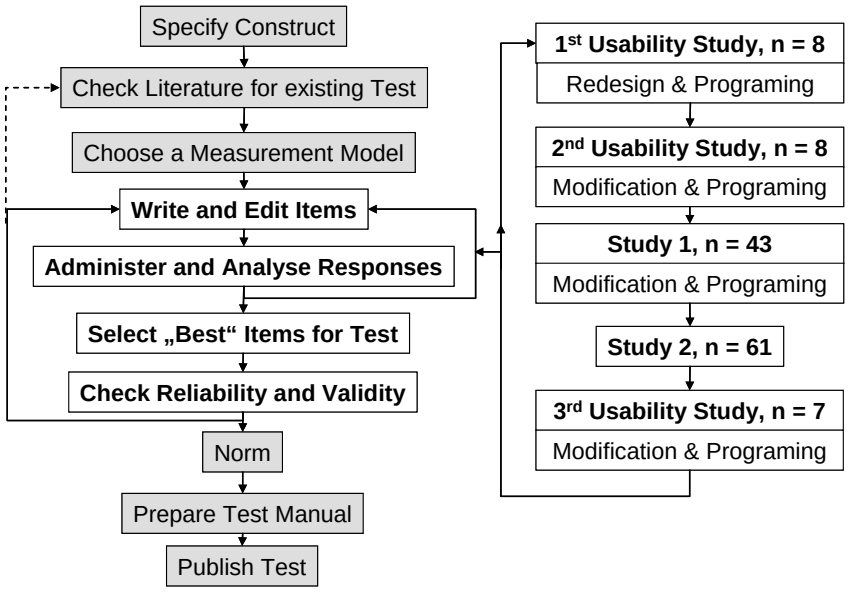


Figure 3: Adapted test development process of the GL based on the traditional approach presented by Shum, O'Gorman, & Myors (2006)

Acceptance and usability of the Genetics Lab among digital natives

A first analysis of our samples' characteristics showed that the claims made by the literature about today's students were well supported by our studies. Figure 4 presents several ICT-related characteristics of the participating students. As can be seen, the vast majority of students is using personal computers already longer than 3 years (92%) and nearly everyday (up to 80%). Moreover, these students report a high ICT-competence on a 10 item questionnaire including several ICT-related activities like burning CDs, downloading pictures and programs from the internet, creating a webpage or a multimedia presentation. Total scores of this scale were (linearly) transformed into percentage of a maximum possible score that could be attained (POMP, see Cohen, Cohen, Aiken, & West, 1999). Thus, the percentages depicted in the black bars of Figure 4 (75% for Study 1 and 78% for Study 2) describe the mean achieved percentages of a maximum achievable ICT-competence score.

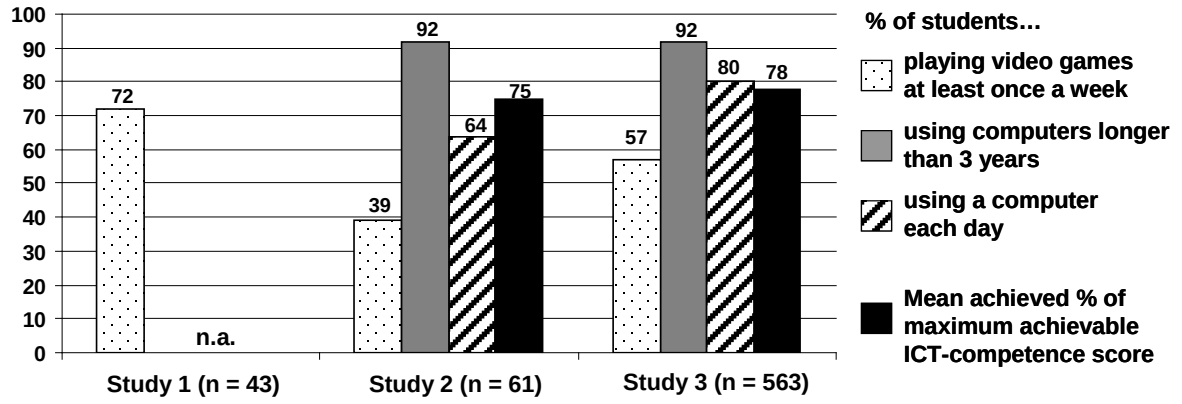


Figure 4: ICT characteristics of students (n.a. is not applicable)

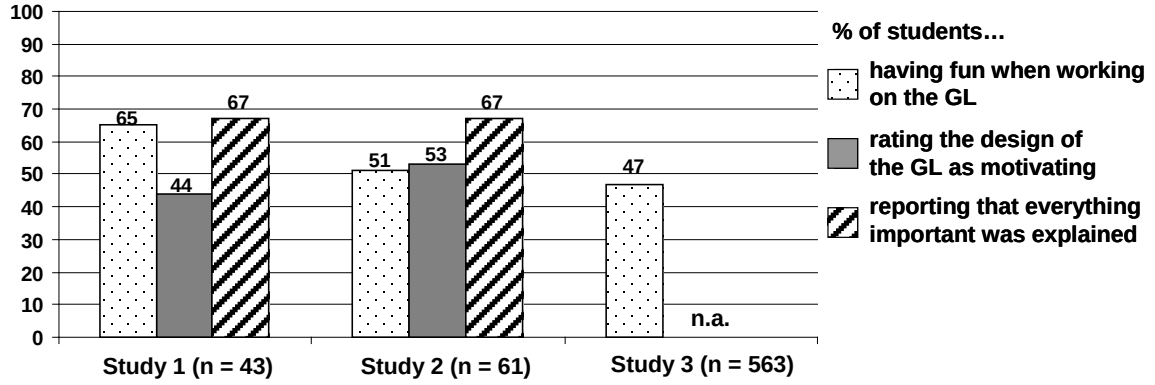


Figure 5: Acceptance of the Genetics Lab among students (n.a. is not applicable)

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

To investigate whether our attempt to develop a microworld suited for today's students was successful, we evaluated acceptance of the GL within the conceptual framework of well-established technology acceptance models (e.g. Terzis & Economides, 2011). According to this

framework, acceptance of a computer-based assessment instrument may be substantially influenced by its *Perceived Ease of Use* and *Attractivity*. Moreover, the instrument's *Comprehensibility*, and *Functionality* are considered as crucial factors determining its usability and hence its acceptance among the target population. Consequently, students participating in Study 1 had to rate various elements of the GL in terms of these four dimensions. Results of this questionnaire are presented as POMP-scores in Table 2. Given the lack of comparable studies, we considered values above 50% - indicating that positive student evaluations outweigh negative evaluations - as positive outcomes (for more details about the questionnaire, please refer to (Sonnleitner, et al., 2012). Overall, results show that the GL is well accepted among today's students. The GL was rated as being easy to use (M = 54, SD = 23) with an attractive appearance (M = 64, SD = 22). Students also attested the GL to be comprehensible (M = 61, SD = 17) and well functioning (M = 60, SD = 22). Moreover, in all three studies, large portions of the students reported having fun while working on the GL (see Fig. 5). Apart from Study 1 in which the GL was solitary administered, students had to work on the GL at the end of a 2 hours test session. This may explain the somewhat smaller portion of students in studies 2 and 3 that indicated to have fun while working on the GL. Moreover, results show that the game-like design of the GL was appreciated and even described as motivating by large portions of the students. Crucially, the vast majority of students felt that everything important was explained during instructions. We take this as clear indication of the instruction's efficiency to successfully illustrate the handling of the GL in an interactive multimedia-based way.

To sum up, results suggest that our attempt to develop a microworld of high usability that enjoys high acceptance among today's students was successful. For the first time, we could also go beyond anecdotic evidence that interacting with such scenarios makes fun (e.g. Ridgway & McCusker, 2003). We largely attribute this positive outcome to the actions we have taken to

consider special characteristics of today's students, namely, the integration of game-like characteristics, the development of a standardized and interactive multimedia instruction and extensive usability testing.

Challenge 2 - Scoring

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

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

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

Students' exploration behaviour

In order to gather knowledge about the effects of a creature's genes on its characteristics, students have to explore it by actively manipulating the genes (see Figure 1, a). These manipulations, however, are most informative if students switch one gene to "on" and all other genes to "off". Only then can occurring changes on the creature's characteristics be unanimously interpreted as effects of the gene that is switched "on" (Vollmeyer, Burns, & Holyoak, 1996). Moreover, in order to detect dynamics within a creature (i.e. some characteristics change without being affected by genes), all genes have to be switched "off". Thus, when looking at the students' traces, such informative steps can be distinguished from non-informative ones. This behavioural information can then be used to relate the number of informative steps to the total number of steps taken in the exploration phase across all creatures, resulting in a behaviour-based *Systematic Exploration* score (Kröner et al., 2005). A high proportion of informative steps thus indicate a very efficient exploration strategy of a student – he mostly applies informative steps.

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

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

Students' gathered knowledge

The students' gathered knowledge about a creature was scored on basis of the causal diagrams which were depicted by the students in the GL's database (see Figure 1, b). These causal models can be interpreted as the theoretical model a student has developed about a creature and are thus valid indicators of his mental problem representations (Funke, 1992). Although the method of knowledge assessment by causal diagrams was often successfully used in samples of university students (Blech & Funke, 2005; Funke, 2001), critique was raised that due to its high cognitive demand, it might be problematic in samples of lower cognitive capacity. This notion, in fact, was supported by the results of our usability studies, showing that many students of our target population reported problems when using causal diagrams for knowledge representation. Although, we tackled this problem by a substantial modification of the related instructions it still had to be confirmed that the analysis of causal diagrams yields reliable and

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

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

Students' control performance

For scoring student's ability to apply the gathered knowledge and achieve certain target values on the creatures' characteristics (see Fig. 2, c), we again drew on behavioural data to

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

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

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

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

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

Summary and Outlook

It has been shown that the assessment of complex problem solving "in the classroom" poses special demands on the assessment instruments used for this purpose. The development of the Genetics Lab successfully responded to most of these challenges by drawing on game-like characteristics, a user interface of high usability, and psychometric sound, behaviour-based scores that are at the same time comprehensive for educators. However, several questions remain to be answered. First, although the vast majority of students in our studies showed characteristics

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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