Flexible Automated Assessment in 3D Learning Environments: Technical Improvements and Expert Feedback

Joachim MADERER[a] and Christian GÜTL[a, b, 1]
[a]*Graz University of Technology, Graz, Austria*
[b]*Curtin University, Perth, Western Australia*.

Abstract. Immersive 3D learning environments have a great potential to improve learning. This holds especially true for the usage of computer simulations in science education. Nevertheless, the importance of formative assessment, guidance and feedback requires the implementation of automated and flexible assessment solutions within these environments. Our workgroup has recently introduced a flexible assessment framework with focus on behavioral assessment and immediate feedback provision. The aim of this paper is to report about the latest technical challenges and improvements of our Open Wonderland prototypes, as well as a first expert evaluation that yields promising results. From the technical perspective, issues with synchronization and the amount of data being transferred have been identified and a first solution has been implemented. The expert evaluation confirms the assessment concept and reveals interest among (emerging) practitioners; however, several ideas to improve the feedback provision and usage of simulations have been stated as well.

Keywords. Immersive environments, STEM education, assessment, science, physics, feedback, guidance, learning, framework, open wonderland, expert evaluation

Introduction

The increasing support of new technologies has enabled the average user to benefit from Internet technologies and rich computer graphics. Certain pedagogical benefits, such as collaborative and contextual learning is ascribed to *3D learning environments*. These include games, simulations and virtual worlds. [1] The latter combine the constructivistic affordances of social interaction and immersive graphics [2].

Given the challenges STEM education is facing [3, 4], these technologies are considered a powerful learning environment [5]. Literature reveals significant evidence that computer simulations improve conceptual understanding and enable *scientific discovery learning (SDL)*. That means students can simulate an entire research cycle, i.e. finding research questions and hypothesis as well as conducting experiment, evaluation and interpretation. Nevertheless, it is also confirmed that guidance is important as an entirely unstructured simulation does hardly foster learning. Besides that, supportive and procedural information is required by students in order to

[1] Corresponding Author.

successfully complete the tasks. But it was found that collaboration also significantly contributes to the learning outcomes. [6] In addition, particularly formative assessment and feedback are important for the success of learning [7].

Literature reveals first approaches for automated assessment in virtual worlds. But there are still several issues. Connections to existing e-learning systems, such as Moodle can at this time only deliver traditional e-learning content and question items, such as multiple-choice questions or predefined conversation patterns [8-10]. There are more advanced and interactive solutions, enabled through in-world scripting. But teachers might not have the skills (and probably time) to script custom assessment solutions. Despite that, such solutions are usually not flexible as they are written for a specific platform. [10] Little has been reported about approaches that would consider complex player behavior for assessment measurements (*cf*. [10]); although, Ibáñez et al. [11] promote the potential as player actions could be easily intercepted. Nevertheless, first examples exist in the context of game-based learning (*see* [12-15]).

Based on similar issues we have recently reported about a conceptual solution architecture for a flexible assessment framework that externalizes the assessment process, and thus supports a variety of platforms and environments. It is based on a semantic-enabled approach and is supposed to target different 3D learning environments. For this approach, a software component, called *assessment module* is a necessary pre-condition that must be implemented once for a certain platform. Thus, our first steps included the development of prototypes for the virtual world platform Open Wonderland (OWL). These prototypes include an assessment module as well as a simulation of a physics experiment that conforms to the approach. [16]

In this paper we briefly report about additional technical challenges and improvements of the prototypes. Based on that, an enhanced showcase is presented. Furthermore, the entire set of implemented showcases including the recently implemented one, was demonstrated to a group of experts, including young teacher trainees (students) but also an instructional expert (school teacher and university lectureship), research expert and two university lecturers. The most important results of this evaluation will be presented and discussed as well.

1. Improvement of the Prototypes

1.1. System Overview

The flexible assessment framework consists of three tiers, including the *immersive software platform*; and arbitrary *assessment system* that contains the actual assessment logic and acts as middleware; as well as possible connections to *external systems* to access learner specific settings and data. The immersive environment and the assessment system communicate via a web service API. The central component is an *assessment module* that is attached to a specific immersive environment and is responsible to compile and forward events to the external assessment system. These events consist of intercepted user interactions and environmental conditions, rather represented on a semantic level instead of raw information. The module is further supposed to process and display incoming feedback messages appropriately on the clients of the participating learners.

In order to adopt the approach for other platforms, the assessment module is supposed to implement the following three dimensions of event construction:

1. *Common events*: Simple user actions which belong to virtual worlds in general – such as moving the avatar or using gestures – should be intercepted in a generic fashion and compiled into "semantic events".
2. *Tagging and metadata components*: The position of the users' avatars should be monitored through the introduction of annotated spatial sections that could also be nested. Appropriate events are generated when an avatar enters or leaves such sections. Besides, all virtual objects should be enhanced with general metadata that identifies the object on a semantic level, e.g. attributes for object classification. Additionally, this also includes the definition of proximity ranges that declare discrete distances to the avatar. This should at least include the general perception of virtual objects as well as an appropriate operational distance.
3. *Programmatically invoked events*: To support more concrete interactions with objects, the most important part is the enhancement of individual object types. This means that a virtual object is generally supposed to report about changes of state as well as object-related interactions of users.

In addition, feedback mechanisms for each type of generalized feedback should be implemented. A web service connection is consequently supposed to deliver the events to an external assessment system and receive feedback commands to be realized through different feedback plugins. The latter could also be implemented as part of the assessment module, for instance, a simple text-based display feature. A more detailed explanation can be found elsewhere [16].

We believe this approach will be flexible in general because semantically self-descriptive objects could be reused in similar settings; and 3D environments have potential to become semantically self-descriptive in the near future. But it appears at this time a commonly accepted standard for such semantics is still outstanding. [17] Furthermore, also Schmeil et al. [18] discuss the relevance of semantic considerations for collaboration in virtual worlds from a conceptual perspective.

1.2. Technical Challenges and Solution Approaches

Two major issues emerged during the development of the prototypes, especially regarding the enhanced showcase that will be introduced in the next subsection. The following paragraphs will briefly explain these issues and sketch their solution.

First, the amount of data being transferred to the external assessment system was a problem. While autonomous state changes of any virtual object should only appear from time to time, real-time simulations consist of rapidly changing properties that can become required for the assessment process. Hence, it was decided to introduce three different levels of data that cumulatively contain each other and define the state of an object or entity:

- *Dynamic* state refers to continuously changing properties. This type of data is only reported together with a user interaction. But the assessment module takes care that each object that is currently in the range of the learner's perception will report its state, independent of which object was involved in the user interaction;
- *Changeable* data includes dynamic data but is extended by information that does not change too often. This level of information is only reported for objects that have directly changed through a user interaction;

- *Full* data updates, which contain also the identification (metadata) of an object, are only included if a learner enters into the range of an object.

Second, due to the potential collaboration between users it is also important to synchronize the simulation model of all clients and servers in the range of a few milliseconds. Because simulations should be rendered as fluently as possible each node is responsible on its own to propagate the simulation. Therefore, it was necessary to negotiate exact time codes between clients and server. This enables also transport delays to be incorporated when the simulation model is updated from time to time in order to prevent an accumulating divergence. However, further findings indicate also that times between user interactions and compilation of event data on the server-side is crucial for an exact representation of the simulation data in the context of the external assessment system. This has raised our interest in a possible support framework that allows for synchronization of arbitrary simulations in OWL based on exact time codes.

1.3. Enhanced Showcase

Based on the improvements that have been discussed in the previous subsection an enhanced showcase could be developed, representing an actual experimental task. Besides the simulation of a *simple pendulum* that has been used for the first show cases [16], an additional object has been added to the context. This *'assignment object'* – depicted as a rotating box with question marks – opens a control panel that contains a *stop watch* as well as an *input field*. The learner is supposed to use the stop watch to measure the periodic time and calculate the current frequency of the pendulum. It is important to note that these two objects – pendulum simulation and stop watch – are technically decoupled objects. That means pressing a button on the stop watch will not consider the state of the pendulum explicitly. However, all other assessment-compliant 3D objects in range – and this includes the pendulum – will be triggered to report their *dynamic state* at the same time.

The external assessment logic separates between two cases. If the input of the frequency was wrong, but the previous measurement appeared to be accurate it suggests that the learner should check his or her calculation. Otherwise, the system recommends repeating the measurement (see Figure 1). This is achieved by comparing the deflection angels of the pendulum at the time user interactions – *starting* and *stopping* the stop watch – are sent to the system. The result of the last measurement is remembered in the context of the assessment system and used when the learner *submits* the calculated frequency.

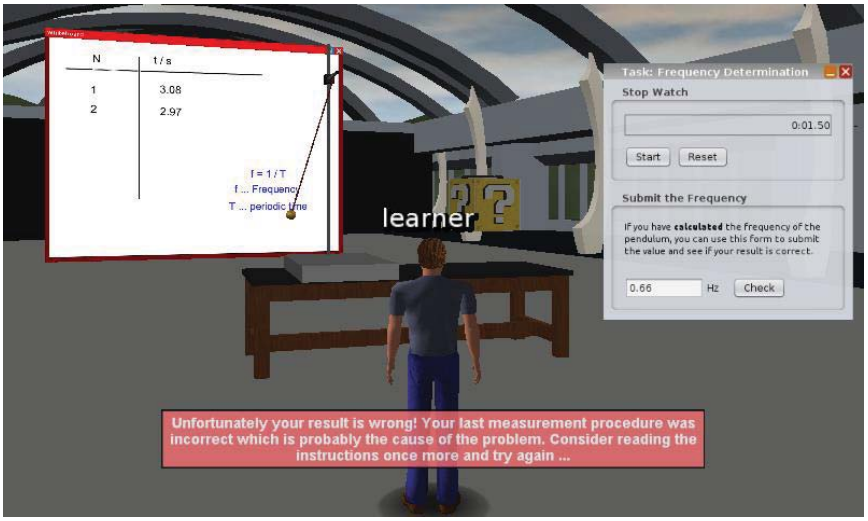

Figure 1. Learner has entered a wrong frequency and the assessment system leads this back to an imprecise or conceptually flawed measurement.

2. Expert Evaluation

The prototypes and test scenarios are not yet completed enough to be used in a real student context. But it seemed necessary at this stage to obtain initial feedback from people who might actually use these systems for teaching. The purpose was to avoid a fundamentally wrong direction, collect suggestions for improvements and to explore which options might be relevant for teachers in order to obtain information about students.

2.1. Methodology

The expert evaluation was conducted as informal, semi structured interviews based on a predefined questionnaire, whereas specific sets of fixed-response questions have been ask prior to the demonstration. The intention was to align the attitudes regarding formative assessment, feedback and the potential of 3D learning environments – but also the individual foreknowledge about e-learning and immersive education – with the results of the actual prototype evaluation.

The demonstration itself included three showcases which have been presented to the experts; whereas the first and second have already been reported in [16] and the third in the previous section:

1. The learner approaches the experiment workplace. Immediate feedback is triggered based on the location which directs the learner towards an in-world PDF reader, providing written instructions;
2. If the learner deflects the pendulum too much, a warning appears, as this would leave the idealized measurement range;

3. The learner is supposed to measure the periodic time of the pendulum, subsequently calculate the frequency, and finally, confirm the result through a submission form. The external assessment system observes the entire process and can provide different feedback. In addition, assessment models have been switched manually in order to demonstrate the difference between more and less intensive feedback provision. That means intermediate feedback is either provided before the final submission or not.

Most of the subsequent questions were also based on a fixed range of answers (e.g. strongly agree to strongly disagree) referring to a Likert scale as well as similar rating scales. The participants were, however, also invited to contribute additional comments and keywords in written form. Besides, particularly important statements and attitudes, if not covered through the questionnaire, have been recorded during the entire session in form of written keywords from the perspective of the interviewer.

2.2. Results

In total nine subjects participated in the study: five teacher trainees in physics (students; age group 20-29, except one 30-39; one female only); a high school teacher in physics who is also university lecturer in subject didactics (male; 50-59); a university lecturer/assistant in experimental physics (male, 30-39), a university lecturer/assistant in chemistry who is also teacher trainee in school physics (female, 20-29); as well as a computer science expert with a research background in immersive education (male, 40-49).

Because all participants of this evaluation are German native speakers, some statements have not been provided in English. All quotations used in this paper are either close translations or might have experienced marginal linguistic improvements.

The *initial* questions[2] revealed that the average do not feel themselves particularly experienced with e-learning (M = 3.11, SD = 0.99; between *almost unfamiliar* and *very experienced*). But almost all participants highly agreed on the importance of formative assessment (M = 4.11, SD = 0.74; between *irrelevant* and *very important*), timely (immediate) feedback (M = 4.56, SD = 0.50), as well as on adapted feedback for individual students or groups (M = 4.44, SD = 0.50). However, besides multiple-choice questions (selected 8 times), novel e-assessment practices appear rather unknown to the participants, as hardly anyone has used e-assessment tests (or could name anything) that goes beyond fixed response questions, numeric or free text answers (approx. selected 4-5 times each). The question if multiple-choice questions are sufficient to evaluate the learning outcomes of science education rather diverges (M = 2.89, SD = 0.87). Some commended that it depends whether it is about factual knowledge or skills and competencies. Nevertheless, experiences with computer games (M = 2.22, SD = 0.63; personally playing, between *never tried out* and *very often*), as well as computer simulations (M = 2.78, SD = 0.79) and 3D virtual worlds (M = 1.89 of 4.00, SD = 0.99; between *never heard* and *professionally used*) for learning activities are practically not existent (the research expert certainly used it professionally). Although the greater part was convinced that 3D learning environments can motivate but also improve learning (M = 3.78, SD = 0.63).

[2] The range of all answers is between (1) and (5), further between *strongly disagree* and *strongly agree*, if not otherwise indicated.

The major *results of the actual prototype evaluation* are listed in Table 1. Results reflect the quite good perception of the prototype and its showcases by the experiment group.

Table 1. Results of the actual prototype evaluation

Question or Statement	Mean (SD)
What is your overall impression[3] of the demonstrated assessment and feedback aspects?	4.11 (0.74)
I think the example is authentic.	4.33 (0.47)
I think the textual feedback provided at the bottom of the window was helpful.	4.44 (0.68)
I think the feedback provided would improve the outcomes/results of students.	4.33 (0.47)
I think the feedback provided would improve the understanding of students.	4.00 (0.67)
I think the kind of player actions evaluated – measurement activity and calculation – can be used in accordance with competency-based learning models – i.e. the approach is valid to reflect on skills and competency levels of the learners.	3.94 (0.68)
I think the different intensity of feedback messages is appropriate to catch up with the different competency levels of students.	4.22 (0.63)

Positive comments included *"well done"*, praised the immediate and individual feedback, as well as the *"challenging tasks as motivation for students"*, and referred to the approach as *conceptually* very good and *interesting*. It was further stated that it was a *"practical experiment"* which is *"easy to handle"*, also including the idea of a PDF containing written instructions. One participant stated that the *"different colors for positive, negative feedback are fine for 'visual types'"*.

Nevertheless, regarding the *negative comments*, one participant contrarily stated that the *"feedback is very generic"* and it is *"difficult to provide feedback individually"*. Other participants suggested *"to force students to read the instructions (e.g. by implementing a control task)"* and that the *assignment box* would not really feel authentic. It should also be mentioned that the quality of the overall surrounding environment was criticized, although that is not directly related to the evaluation of the assessment concept. Particularly one participant, who was generally less fond of computer-based and 3D virtual world approaches stated: *"3D graphics is in my view not necessary required (with the pendulum)"*. In addition, also the feedback was considered too small and positive feedback could still be displayed brighter. Especially two participants who were less convinced on the benefits of 3D learning environments had a hard time to focus on the assessment concept and concentrated a lot on general imperfections that arose from the used platform, materials and exemplary approach.

Several (easier to implement) *improvements* can directly be extracted from this feedback but there have further been more explicit recommendations, concerns that implicitly validate the requirement for this approach, as well as further ideas for future developments:

- Sound should be added, and maybe also *"laboratory music"*;

[3] Question is rated between insufficient (1) and very good (5); all other statements between strongly disagree (1) and strongly agree (5).

- An important recommendation was that the computer should read the feedback aloud*;*
- One participant stated that *"more comments and hints regarding the expected actions would be fine"*, he or she guesses that *"students who are not that talented could be disappointed since they might have problems with starting their own exploration of the virtual world"*;
- Movements should be combined with numerical representations;
- A pocket calculator should be added in-world;
- Minor aspects as the position of feedback could be improved;
- To facilitate a game-based approach, for instance, the explanation of a formula could be released as a reward for achievements in the practical exercise;
- The considerations of external influences, in the context of a pendulum simulation for instance an eddy current brake.

In addition, several questions were dedicated to *decisions for future developments* and an integrated stack of systems, including feedback for teachers. The greater part has expressed interest for information at a glance, including overview of students' problem domains (selected 9 times), as well as an overview on the entire classroom or groups (7 times). One participant added an additional item and expressed interest on the collaboration activities of students.

Besides that, the participants were asked if a challenging aspect, e.g. progress information among fellow students or groups, might improve the motivation for learning (*cf*. [19]). Most experts agreed with that (M = 4.44, SD = 0.68). When it comes to the incorporation of assessment information from such virtual activities into the grading schema, the answers are less clear and have a larger divergence again (M = 3.78, SD = 0.92). At least one of the participants seemed to be concerned about legal issues. Nevertheless, almost all participants agreed positively on the idea to offer a graphical editor system to design assessment rules on their own (M = 4.00, SD = 0.67); although, some did not feel themselves particularly capable of basic programming during the pre-questionnaire. In addition, it is also worth mentioning that during the informal interview process it became clear, that some participants were concerned on the available time for both, review on student information as well as the design of assessment rules.

Finally, the motivation to use immersive 3D virtual worlds in different application contexts was acquired (see Table 2). Most experts clearly considered it a supportive measurement for additional exercises when conventional material is used without practical experiments. Other usage options included the comparison between model and reality, homework exercises, as well as additional exercises following practical experiments. In addition, further value on the application of 3D virtual worlds was seen regarding communicative aspects, game-based environments, training, concept explanation, the reduced necessity to read, as well as experiments which are complicated or not possible to be implemented in real world settings.

Table 2. Types of application of immersive 3D virtual worlds in physics education (predefined categories)

Usage option	Count
Preparation for real practical lessons	4
As support for courses/activities which do not feature practical (laboratory) lessons	9

If embedded in a greater context, as replacement for real practical activities	3
Not at all (please provide comment)	0
Other usage	5

3. Discussion and Outlook

The aim of this paper was to report about the latest findings of a flexible assessment framework that is able to support different application domains and immersive learning platforms based on a semantic-enabled approach. The first prototype of an assessment module has been implemented in Open Wonderland in the context of STEM education; whereas a simulation of a simple pendulum was supported through external feedback messages to guide the learner (*see* [16]).

The first section was concerned with technical challenges and solution approaches that occurred during the ongoing development. Especially two critical aspects have been examined. *First*, it was necessary to introduce different levels of semantic-enabled data to prevent an overstress of the communication layer and external assessment service. Continuously fluctuating values of simulations will only be reported based on context and related user interaction. *Second*, in order to provide fluent simulations among different clients, and to provide accurate information for server-side and external assessment systems, it was necessary to reliably synchronize clients and server based on exact time codes. This will need further research and improvement in the future. Additionally, an enhanced showcase was introduced that depends on this synchronization and allows learners to determine the periodic time and frequency of the swinging pendulum.

The second section reported about the methodology and results of a first expert evaluation. Three showcases were demonstrated to a group of nine experts, consisting of teacher trainees in physics, practicing teachers and related research experts. To subsume, the greater part of the participants was quite interested, although not particularly aware of e-learning approaches (electronic assessment) and especially immersive 3D environments. Nevertheless, most experts would welcome such integrated tools as part of an available e-learning solution, thus confirming the overall concept. Besides that, the need for guidance and individual feedback is significantly confirmed (*cf*. [6], [7]), which would also justify the need for such a flexible approach in general. Negative attitudes towards computer-supported education, more precisely 3D learning environments, also matched with a less euphoric evaluation of the prototype which is less surprising. Beyond that, several recommendations have been given for further improvements. Some of them can be realized rather easily, others require more afford, such as spoken feedback; but the latter not less interesting, considering the basic idea of an immersive computer environment.

Based on these findings, we consider it promising to further investigate this approach on different dimensions. The next steps should include a proper implementation of an enclosed learning setting and let students experiment with the scenario. Furthermore, the coupling of the assessment module with non-player characters (*cf*. [11], [16]) offers potential for an additional feedback mechanism in the near future, which might even uses a speech synthesizer to provide also an auditive source of feedback. Regarding the provision of real individual feedback, it is still open to connect the external assessment system with learning management systems to access

preferences and learner profiles to provide custom feedback. Another aspect refers to the usage of different assessment logic, which – in contrast to a simple rule-based assessment engine – might rather be based on more advanced solutions in artificial intelligence. These issues should determine the next series of research projects.

Acknowledgements

Special thanks to Dr. Mohammad AL-Smadi, who was part of the research team and initial contributor to service-oriented flexible assessment approaches. Further appreciation is dedicated to Dr. Gudrun Wesiak for her recommendations. The first steps of this research have been supported by the European Commission as part of the ALICE project (http://www.aliceproject.eu).

We are further grateful to the experts who made their time available to participate in the prototype evaluation and provided valuable feedback.

This paper is also linked to aspects of the research project nDIVE (http://ndive-project.com). Support for the production of this publication has been partly provided by the Australian Government Office for Learning and Teaching (Development of an authentic training environment to support skill acquisition in logistics & supply chain management, ID:ID12-2498). The views expressed in this publication do not necessarily reflect the views of the Australian Government Office for Learning and Teaching.

References

[1] B. Dalgarno, M. J. Lee, What are the learning affordances of 3-D virtual environments?, *British Journal of Educational Technology* **41**, 10–32 (2010).

[2] M. D. Dickey, Teaching in 3D: Pedagogical affordances and constraints of 3D virtual worlds for synchronous distance learning, *Distance education* **24**, 105–121 (2003).

[3] J. Osborne, J. Dillon, "Science education in Europe: Critical reflections" (Nuffield Foundation, London, 2008).

[4] R. W. Bybee, What is STEM education?, *Science* **329**, 996–996 (2010).

[5] T. Machet, D. Lowe, C. Gütl, On the potential for using immersive virtual environments to support laboratory experiment contextualisation, *European Journal of Engineering Education* **37**, 527–540 (2012).

[6] N. Rutten, W. R. van Joolingen, J. T. van der Veen, The learning effects of computer simulations in science education, *Computers & Education* **58**, 136–153 (2012).

[7] D. J. Nicol, D. Macfarlane-Dick, Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, *Studies in higher education* **31**, 199–218 (2006).

[8] D. M. Arroyo et al., Assessment in 3D virtual worlds: QTI in Wonderland, *Congreso Iberoamericano de Informática Educativa, IE2010*, Santiago de Chile, Chile, 2010.

[9] G. Crisp, M. Hillier, S. Joarder, Assessing students in Second Life – some options, *Curriculum, technology & transformation for an unknown future*. *Proceedings ASCILITE Sydney*, 2010, 256–261.

[10] G. Crisp, Meaningful assessment within virtual worlds – what should it look like?, 2012, Retrieved from http://www.inter-disciplinary.net/wp-content/uploads/2012/02/crispepaper.pdf.

[11] M. B. Ibáñez, R. M. Crespo, C. D. Kloos, in *Key Competencies in the Knowledge Society*, N. Reynolds, M. Turcsányi-Szabó, Eds. (Springer, 2010), pp. 165–176.

[12] V. J. Shute, M. Ventura, M. Bauer, D. Zapata-Rivera, in *Serious games: Mechanisms and effects*, U. Ritterfeld, M. Cody, P. Vorderer, Eds. (Routledge, 2009), pp. 295–321.

[13] M. D. Kickmeier-Rust, C. M. Steiner, D. Albert, in *Intelligent Networking and Collaborative Systems, 2009. INCOS'09. International Conference on*, (IEEE, 2009), pp. 301–305.

[14] M. D. Kickmeier-Rust, D. Albert, Micro-adaptivity: protecting immersion in didactically adaptive digital educational games, *Journal of Computer Assisted Learning* **26**, 95–105 (2010).

[15] M. AL-Smadi, G. Wesiak, C. Gütl, in *Interactive Collaborative Learning (ICL), 2012 15th International Conference on*, (2012), pp. 1–6.
[16] J. Maderer, C. Gütl, M. AL-Smadi, Formative Assessment in Immersive Environments: A Semantic Approach to Automated Evaluation of User Behavior in Open Wonderland, *Journal of Immersive Education – Proceedings of iED 2013 Boston Summit*, *Boston*, 2013.
[17] T. Tutenel, R. Bidarra, R. M. Smelik, K. J. D. Kraker, The role of semantics in games and simulations, *Computers in Entertainment* **6**, 57:1–57:35 (2008).
[18] A. Schmeil, M. J. Eppler, S. de Freitas, in *Engaging the Avatar*. *New Frontiers in Immersive Education*, R. Hinrichs, C. Wankel, Eds. (Information Age Publishing, 2012), pp. 15–48.
[19] M. Prensky, "Simulations: Are They Games?", in *Digital Game-Based Learning* (McGraw-Hill, 2001). Available from: http://www.marcprensky.com/writing/Prensky%20-%20Simulations-Are%20They%20Games.pdf