Accommodating Stealth Assessment in Serious Games: Towards Developing A Generic Tool

Konstantinos Georgiadis
Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences
Open University of the Netherlands
Heerlen, The Netherlands
konstantinos.georgiadis@ou.nl

Giel van Lankveld
Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences
Open University of the Netherlands
Heerlen, The Netherlands
giel.vanlankveld@ou.nl

Kiavash Bahreini
Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences
Open University of the Netherlands
Heerlen, The Netherlands
kiavash.bahreini@ou.nl

Wim Westera
Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences
Open University of the Netherlands
Heerlen, The Netherlands
wim.westera@ou.nl

***Abstract*—Stealth assessment derives the progression of learning in an unobtrusive way from observed gameplay captured in log files. To this end, it uses machine learning technologies to provide probabilistic reasoning over established latent competency variable models. Now that video games are increasingly being used for training and learning purposes, stealth assessment could provide an excellent means of monitoring learning progress without the need for explicit testing. However, applying stealth assessment is a complex and laborious process. This paper analyses the limitations of stealth assessment and conceptualizes the requirements for developing a generic tool that could overcome its barriers and accommodate its practical application. Hence, a framework is presented describing its user and functional requirements. The proposed generic solution could open up the wider uptake of stealth assessment in serious games.**

Keywords—stealth assessment, machine learning, serious games, learning, generic tool

I. Introduction

Serious games intend to support learning in education and training [1]. In recent years, serious games have gained attention [2] for their potential to enable active modes of learning for the acquisition of knowledge and skills [3]. However, more research is needed in developing game design methodologies and tools to enforce their pedagogical outcomes [4]; because games are inherently complex by nature and usually characterized by fuzzy game design practices driven from creativity rather than pedagogy [5]. This often leads to disputes regarding their educational value [6]. Hence, their embodiment in educational settings is still scarce [7].

Nevertheless, a growing body of studies is attempting to address these issues. Most notably, an emphasis is given to the importance of formative assessments in serious games [8]. One way to achieve this is to base these assessments around valid and reliable competency constructs [9]. The term competency is a fuzzy concept usually meaning a set of knowledge, skills, and abilities [10].

One of the most prominent methodologies for formative assessments in games is referred to as Stealth Assessment (SA) [11]. SA is an unobtrusive methodology directly integrated in the gaming environment, which exploits emerging data from gameplay to computationally analyse and interpret learners' performance through machine learning (ML) technologies. Being unobtrusively embedded in games, SA does not disrupt the learning experience thus allowing the learners to enter into a "flow" state which is regarded to be conductive to engagement, hence also to learning [12]. In addition, it reduces the saliency of the assessment process, which minimizes the test anxiety and improves the validity of the assessment itself [13]. In contrast to traditional test items (e.g. self-report questionnaires, multiple-choice tests, etc.), it can (a) benefit from the opportunity to access high resolution data, thus allowing for more detailed assessment of the learners, (b) allow the adaptation of the game's instructional design to meet the learner's personal needs, (c) counter inherent issues in traditional tests such as social desirability effects, and (d) provide rapid feedback during the learning process.

Georgiadis, K., van Lankveld, G., Bahreini, K. & Westera, W. (2018). Accommodating Stealth Assessment in Serious Games: Towards Developing A Generic Tool, 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games Conference Oct 23, 2018). Würzburg, Germany: IEEE, p. 1-4.

However, SA encounters various limitations, especially regarding its applicability. It is acknowledged that the implementation of SA in serious games is a labour intensive, complex, and time-consuming process [13] that requires a broad range of expertise at each step of its realisation as well as open access to source code of existing games and ML tools. Even if the aforementioned requirements are fulfilled, still it is rather intuitive how one can properly map in-game behaviours to competency constructs. Shute and colleagues [14] describe this as an iterative process of brainstorming and pilot testing. Nonetheless, SA also inherits a set of fundamental problems in ML and statistics, such as overfitting [15]. Mainly due to these limitations, SA has not yet been widely adopted by the serious game community.

This study is introducing a conceptual framework for the development of a generic tool, which could accommodate the practical application of SA in serious games. This tool aims to tackle the aforementioned limitations, and thus contribute to the wider uptake of SA in serious games. For this reason, first a description of SA in serious games is presented in Section 2. Next, limitations of SA drawn from relevant literature follow in Section 3. The proposed conceptual framework and the requirements for the development of a generic SA tool can be found in Section 4. Conclusions, remarks, and the future roadmap towards the realisation of the generic SA tool are discussed in Section 5.

II. Stealth Assessment

SA is an unobtrusive evidence-based assessment methodology, which employs ML technologies to provide probabilistic reasoning about learners' performance in serious games. To achieve this, SA utilizes a conceptual framework for establishing relationships between observables from gameplay and competency constructs, which in turn translate to latent variable models that ML algorithms can process. In specific, SA originally combines two main ingredients: (a) the Evidence-Centered Design (ECD) [16] and (b) a ML algorithm called Bayesian Network (BN) [17].

A. The Evidence-Centered Design

ECD is a conceptual assessment framework consisting of three major elements: the competency model, the task model, and the evidence model (see Fig. 1). The competency model defines the construct that describes the underlying factors (i.e. facets) constituting a competency. The task model describes a set of activities in the game that can elicit evidence relating to the competency. The evidence model describes the criteria of how learners' observed performances in-game link to both the competency and task model. The evidence model is described by two components: the evidence rules and the statistical model, respectively. The evidence rules cover the relationship between the tasks and the observed performances, while the statistical model defines the statistical relationships between observed performances and a competency construct. The latter constitutes a latent variable model.

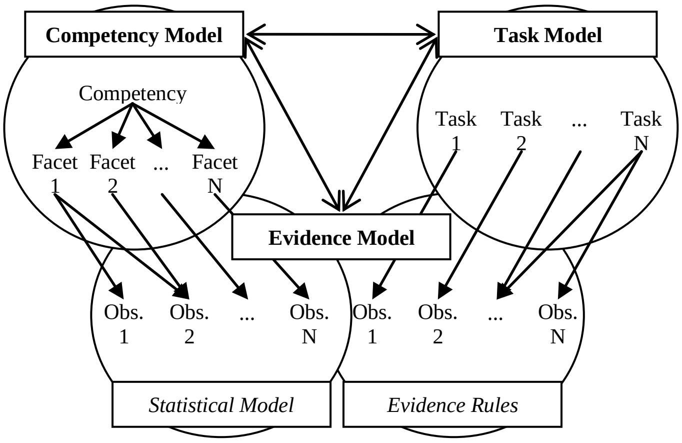

Fig. 1. A view of the Evidence-Centered Design.

B. Machine Learning in SA

SA applies ML to a latent variable model as defined by the statistical model to provide probabilistic reasoning over learners' observed performances. Originally, the use of BNs was considered for this purpose [11]. So far, BNs have been proved to be robust in producing valid and reliable probability statements regarding the mastery of respective competencies. In specific, they have been successfully applied for assessing qualitative physics [18], persistence [19], and problem-solving skills [14]. However, alternative supervised ML algorithms have also been proposed for SA, such as Decision Trees [20].

III. Limitations

This section discusses limitations, such as complexity and laboriousness, in applying SA in serious games.

A. *Complexity*

The complexity of applying SA in serious games relates to the expertise that is needed with account to three perspectives.

Firstly, a certain level of expertise is required from a technical perspective. That is, both game development and ML expertise. Game development expertise mainly refers to knowledge of programming languages and game development tools that allow modifying or developing from scratch a serious game that suits the learning goals. In addition, game (or instructional) design knowledge is needed to properly engineer the in-game tasks by steering the graphical user interface, the narrative, the levels, the audio, etc. This is crucial for eliciting proper evidence with minimum noise introduced from irrelevant covariates that may affect the learning process. ML expertise means knowledge on how to properly implement ML algorithms by taking into account their representation, evaluation, and optimization aspects [21]. It is crucial to underline the importance of transparency regarding those aspects as to allow replication, ensure scientific integrity, and secure pedagogical value.

Secondly, a great amount of expertise is needed from an educational and psychometric perspective. This includes knowledge about the underlying competency constructs, the learning material (to set the game content), and the evidence that maps to the competency constructs (to set the statistical model) and defines the mastery levels (to label the data).

Thirdly, expertise is needed from a statistics perspective. This includes knowledge of statistical methods such as correlation and factor analysis in order to be able to develop, validate, and verify competency constructs. It is essential to underline the importance of validating and verifying the competency constructs as to avoid threats such as construct underrepresentation and construct irrelevant variance [22].

B. *Laboriousness*

Apart from being complex, the process of applying SA in serious games is also quite laborious. SA is originally defined as an assessment methodology that is directly woven into the game environment fabric [11]. As a result, all existing applications have been developed in a hardcoded manner, meaning embedded in the game source code itself. Thus, every time SA is to be applied, it would require software development and validation from scratch. This comes at a great cost, as multiple steps are involved for setting the entire workflow, introducing unattractive routine works that can be prone to mistakes. Even if one manages to successfully apply SA, reconfiguring the system to fit new or updated assessment needs requires additional scripting and manual tweaking on all fronts. The laboriousness of applying SA has contributed in the development of weak business cases so far.

IV. Towards Developing A Generic Stealth Assessment Tool

To tackle the limitations of SA and accommodate its practical application, this study introduces a conceptual framework for the development of a generic tool. Within this framework, SA is described as a stand-alone software tool that is detached from the game source code. In this way, the need for game development expertise to apply SA could be eliminated, while log files from any serious game could be used without the need for additional manual labour.

We argue, that implementing SA as a stand-alone software tool is feasible due to the inherent generic nature of SA. That is because its main ingredients, the ECD and the ML algorithms, can support generic construct representations. In detail, ECD can describe any competency construct within the competency model, while ML algorithms can adjust their representation to match any statistical model; regardless of shape and size. In addition, the number of tasks or observables declared in ECD is not restricted to a specific amount due to a specific threshold. The same holds for the relationships that can be expressed within ECD. Instead, they could be defined on a case-by-case basis following particular assessment needs.

Being part of the RAGE project (rageproject.eu), the development of this generic tool could benefit from the RAGE architecture of applied gaming components [23] to ensure its interoperability and portability. For example, we envision the possibility of hooking up its output with other self-contained software components for providing feedback, adaptation, and learning analytics. For this reason, we consider as a first step the development of a software prototype as a client-side console application. A high order view of the proposed generic framework for SA compared to the original is presented in Fig. 2. The set of requirements for the generic tool follow.

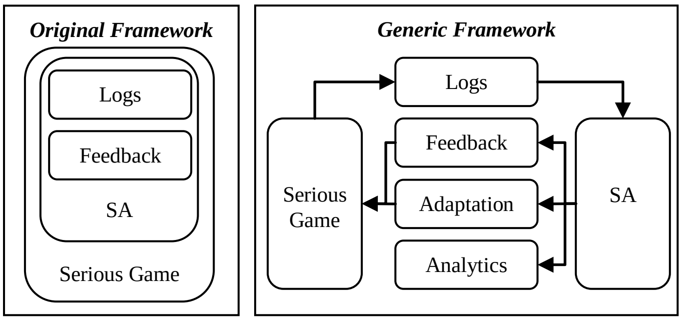

Fig. 2. A view of the original (left) and the generic (right) SA frameworks

A. User Requirements

The users of the generic software tool should not necessarily acquire SA expertise. For example, the users could be game developers, educators, or any other candidate assessor. However, the usability of the tool should be examined to detail the user requirements and accordingly develop support functions and widgets that could reduce the complexities SA pose and enforce the guidance of the users.

B. Functional Requirements

The main focus of the generic tool is to provide functionalities that reduce the amount of expertise and manual labour needed to apply SA. To reduce the amount of expertise needed from an educational and psychometric perspective, the tool should be able to assist its users to easily define their assessment optimizations regarding ECD, game logs, and ML. To minimize the needed for ML expertise, the tool should also provide access to automated built-in ML functions.

Therefore, three functional requirements are set for optimizing the assessment. First, the tool should allow setting the ECD. Second, it should allow importing data from log files. Third, it should allow declaring desirable ML optimizations (e.g. the ML algorithm type and its inner options). It is important to notice that except from supervised ML algorithms, unsupervised ML algorithms should also be available to allow the use of unlabelled datasets.

Moreover, three functional requirements are set for its ML functions. First, the tool should be able to apply different ML algorithms that automatically adjust their representation according to the provided ECD. Second, it should allow the automatic execution of the selected ML algorithms. Third, it should automatically produce detailed output for both students' and ML algorithms' performances for evaluation purposes.

C. Technical Design

To realise these functions, we propose a technical design of the generic tool that includes two main subsystems: (a) a software wizard to tailor the procedure of setting the assessment optimizations, and (b) a machine learning software component to serve the ML functions. A view of the proposed technical design is illustrated in Fig. 3.

D. Support Functions

The proposed technical design can also serve as a means for enabling a set of support functions that could further nurture the process of applying SA. First, a support function could be available to allow the users to validate and verify their competency constructs by automatically applying correlation analysis on the outputs from an external measure (e.g. psychometric test, expert ratings, etc.) and the generic tool. Secondly, as the users of the generic tool may not always be able to define the ECD, a set of support function should be available to assist them. If the users have access to raw data from an external assessment measure (which means that a competency model is also available) and data from a log file, but no knowledge of the statistical model, then a support function could be available to attempt its automatic generation by applying a correlation analysis approach. If the users have only access to a log file but no knowledge of the competency model, then a support function could be available to attempt its automatic generation through a factor analysis approach. A series of forthcoming empirical studies will detail the aforementioned support functions of the generic tool.

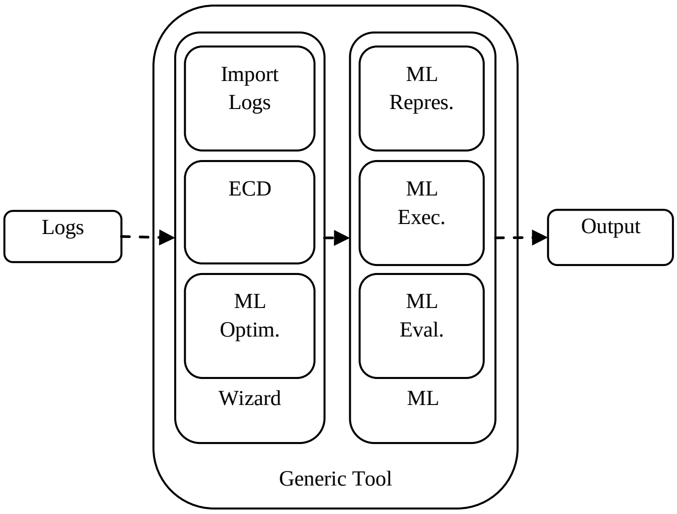

Fig. 3. View of the generic tool for SA in serious games.

V. Discussion

A framework towards the development of a generic solution in the form of a stand-alone software component has been proposed, covering its requirements and additional support functions. We argue that the development of a generic tool within the proposed framework will lead to the wider uptake of SA from the serious game community and also allow reaching a higher level of transparency regarding the educational value of serious games. Despite tackling the major limitations for applying SA, few open issues related to inherent problems in ML and statistics (e.g. overfitting) remain. In addition, we would like to raise awareness on the ethical and social aspects of unobtrusive assessment and thus underline the importance of students' consents.

Nevertheless, the development of a software prototype is already underway and a series of empirical studies will follow to detail and validate the practicability of the generic tool. In the future, we consider the combination of the generic tool with external software to provide feedback, adaptation, and learning analytics. Lastly, we examine its potential application in other domains of education (e.g. MOOCs) or even other scientific fields (e.g. Artificial Intelligence).

Acknowledgment

This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 644187, the RAGE project (www.rageproject.eu).

References

[1] A. De Gloria, F. Bellotti, and R. Berta, "Serious Games for education and training." International Journal of Serious Games, vol. 1, no. 1, 2014.

[2] J. P. Gee, What video games have to teach us about learning and literacy. Macmillan, 2014.

[3] P. Wouters, C. Van Nimwegen, H. Van Oostendorp, and E. Van Der Spek, "A meta-analysis of the cognitive and motivational effects of serious games." Journal of educational psychology, vol. 105, no. 2, May 2013, p. 249.

[4] C. Girard, J. Ecalle, A. Magnan, "Serious games as new educational tools: how effective are they? A meta-analysis of recent studies." Journal of Computer Assisted Learning. vol. 29, no. 3, 2013, pp. 207-219.

[5] W. Westera, "How people learn while playing serious games: A computational modelling approach." Journal of Computational Science, vol. 18, 2017, pp.32-45.

[6] W. Westera, "Games are motivating, aren´t they? Disputing the arguments for digital game-based learning." International Journal of Serious Games, vol. 2, no. 2, 2015.

[7] C. Linehan, B. Kirman, S. Lawson, G. Chan, "Practical, appropriate, empirically-validated guidelines for designing educational games." InProceedings of the SIGCHI conference on human factors in computing systems, 2011, pp. 1979-1988.

[8] V. J. Shute, G. R. Moore. Consistency and validity in game-based stealth assessment. Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective, 2017, pp. 31-51.

[9] B. R. Belland. The role of construct definition in the creation of formative assessments in game-based learning. In Assessment in Game-Based Learning, 2012, pp. 29-42, Springer, New York, NY.

[10] F.D. Le Deist, J. Winterton, "What is competence?" Human resource development international, vol. 8, no. 1, 2005, pp. 27-46.

[11] V. J. Shute. “Stealth assessment in computer-based games to support learning.” Computer games and instruction, vol. 55, no. 2, 2011, pp. 503-524.

[12] V. J. Shute, M. Ventura, M. Bauer, D. Zapata-Rivera. “Melding the power of serious games and embedded assessment to monitor and foster learning.”, Serious games: Mechanisms and effects 2, 2009, pp. 295-321.

[13] G. R. Moore, V. J. Shute, Improving Learning Through Stealth Assessment of Conscientiousness, In Handbook on Digital Learning for K-12 Schools, 2017, pp. 355-368, Springer, Cham.

[14] V. J. Shute, L. Wang, S. Greiff, W. Zhao, G. Moore, “Measuring problem solving skills via stealth assessment in an engaging video game.”, Computers in Human Behavior, vol. 63, 2016, pp. 106-117.

[15] D. M. Hawkins. “The problem of overfitting.”, Journal of chemical information and computer sciences, vol. 44, no. 1, 2004, pp. 1-12.

[16] R. J. Mislevy, L. S. Steinberg, R. G. Almond. “Focus article: On the structure of educational assessments.” Measurement: Interdisciplinary research and perspectives, vol. 1, no. 1, 2003, pp. 3-62.

[17] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier, 2014.

[18] V. J. Shute, M. Ventura, Y. J. Kim. “Assessment and learning of qualitative physics in newton's playground.” The Journal of Educational Research, vol. 106, no. 6, 2013, pp. 423-430.

[19] M. Ventura, V. Shute, M. Small. “Assessing persistence in educational games.” Design recommendations for adaptive intelligent tutoring systems: Learner modeling, vol. 2, 2014, pp. 93-101.

[20] J. L. Sabourin, L. R. Shores, B.W. Mott, J. C. Lester. “Understanding and predicting student self-regulated learning strategies in game-based learning environments.” International Journal of Artificial Intelligence in Education, vol. 23, no. 1-4, 2013, pp. 94-114.

[21] P. Domingos. “A few useful things to know about machine learning.” Communications of the ACM, vol. 55, no. 10, 2012, pp. 78-87.

[22] S. Messick. “Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.” American psychologist, vol. 50, no. 9, 1995, p. 741.

[23] W. Van der Vegt, W. Westera, E. Nyamsuren, A. Georgiev, I. M. Ortiz. “RAGE architecture for reusable serious gaming technology components.” International Journal of Computer Games Technology, 2016, p. 3.