Reinforcing Stealth Assessment in Serious Games

Konstantinos Georgiadis[1[0000-0003-2277-5256]], Giel van Lankveld[2[0000-0001-8319-2244]], Kiavash Bahreini[1[0000-0001-9016-9894]] and Wim Westera[1[0000-0003-2389-3107]]

[1] Open University of the Netherlands, 6419AT Heerlen, The Netherlands
[2] Fontys Applied University of Eindhoven, 5612 AR Eindhoven, The Netherlands
konstantinos.georgiadis@ou.nl
gielvanlankveld@protonmail.com
kiavashbahreini@gmail.com
wim.westera@ou.nl

Abstract. Stealth assessment is a principled assessment methodology proposed for serious games that uses statistical models and machine learning technology to infer players' mastery levels from logged gameplay data. Although stealth assessment has been proven to be valid and reliable, its application is complex, laborious, and time-consuming. A generic stealth assessment tool (GSAT), proven for its robustness with simulation data, has been proposed to resolve these issues. In this study, GSAT's robustness is further investigated by using real-world data collected from a serious game on personality traits and validated with an associated personality questionnaire (NEO PI-R). To achieve this, (a) a stepwise regression approach was followed for generating statistical models from logged data for the big five personality traits (OCEAN model), (b) the statistical models are then used with GSAT to produce inferences regarding learners' mastery level on these personality traits, and (c) the validity of GSAT's outcomes are examined through a correlation analysis using the results of the NEO PI-R questionnaire. Despite the small dataset GSAT was capable of making inferences on players' personality traits. This study has demonstrated the practicable feasibility of the SA methodology with GSAT and provides a showcase for its wider application in serious games.

Keywords: Stealth Assessment, Serious Games, Generic Tool, Statistical Model, Machine Learning, Stepwise Regression, Personality Traits.

1 Introduction

During the last couple of decades the educational community has been putting an increased effort on gradually transcending from traditional classrooms to digital educational environments. These digital learning environments require and can promote specific skills, e.g. critical 21st century skills and abilities, and thus prepare learners for future challenges in workplace and generally in life [1]. Among the most promising forms of digital education are serious games, due to their potential for enabling active learning in rich simulation environments. In these highly dynamic and interactive learning environments it is of vital importance to accurately diagnose the progressing competence level of learners for properly tailoring the learning process.

One of the most promising assessment methodologies proposed for usage in serious games is stealth assessment (SA) [2]. SA combines a principled assessment design framework, namely the Evidence-Centered Design (ECD) [3], with machine learning (ML) technology in order to produce inferences about the learner's competences. The ECD serves as a framework for designing conceptual models for relating competencies (i.e. knowledge, skills, abilities, traits, etc.) and in-game tasks, whilst it also allows for developing computational models that express the relationship of these constructs with evidence (i.e. data) collected during gameplay. These computational models can be processed by ML algorithms and hence produce classifications of the learner's competence levels.

Although it has already been proven in several cases [4, 5, 6] that SA can produce valid and reliable assessments, its practical application is troublesome since it is a complex, laborious, and time-consuming process [7]. SA is inherently complex due to the diversity of expertise that is required in several domains beyond the learning content such as game development and design, machine learning, learning materials, statistics, etc. SA is laborious insofar it has only been applied in a hardcoded manner as an integral part of the games' source code. Such solutions limit the transferability of SA, while it requires each time software development and validation from scratch. As a result, applying SA in a game becomes a time-consuming process that is vulnerable to mistakes.

To overcome the practical drawbacks of SA and accommodate its wider application, a generic solution has been proposed [8, 9]. That is a stand-alone software tool, the Generic Stealth Assessment Tool (GSAT), which (1) allows the use of numerical datasets from any serious game, (2) automates the ML processes, and (3) allows the easy arrangement of different ECD models. GSAT has already been proven for its robustness against simulation datasets [10]. The aim of this study is to examine the use of GSAT with real-world data from a serious game, while concurrently allowing the detailing of the methodology that needs to be followed for this purpose.

To achieve this, data collected in another study [11] from a serious game called *THE POISONED LAKE* is used. This game is intended to allow for capturing behavioural responses that relate to personality traits. These personality traits are described by a five factor model called the OCEAN model [12]. In specific, the personality traits are: (1) Openness to new experiences, (2) Conscientiousness, (3) Extraversion, (4) Agreeableness, and (5) Neuroticism. These traits are also referred to as the "big five personality traits". Apart from collected game data, the study included the data collected from of a valid external measurement for these traits: the NEO PI-R [13] questionnaire. Based on the aforementioned datasets, the authors of the study managed to generate computational models (i.e. statistical models) to relate in-game behaviours with the personality traits, following a stepwise regression analysis method.

In this study, the produced computational models from *THE POISONED LAKE* game are being used with GSAT to directly determine the competence level of the learners on the big five personality traits from the logged player data. The outcomes of GSAT are then compared to the normed scores of the participants from the NEO PI-R questionnaire.

The structure of this paper is as follows. Background information about SA is provided in section 2. Information on GSAT is presented in section 3. Background information on the big five personality traits can be found in section 4. Details regarding the game, the collected data, and the produced computational models are presented in section 5. The methodology that was used for the purposes of the study is described in section 6. Section 7 presents the results of this study, while a discussion over the results and our final conclusions are in section 8.

2 Stealth Assessment Background

As previously mentioned, SA combines the use of the ECD framework with ML technology. These two ingredients are briefly presented in this section.

2.1 Evidence-Centered Design

To arrange assessments in serious games, SA uses a principled assessment design framework called ECD. The ECD consists of several generic conceptual models. In particular, these models are: (1) the competency model, (2) the task model, and (3) the evidence model. The competency model describes the assessed competency as a construct that includes its underlying factors (i.e. facets, sub-skills, etc.). The task model describes a set of in-game tasks that can elicit evidence for the assessed competency. The evidence model allows for describing the relationships of the observed in-game behaviour (i.e. observables or game variables) to both the in-game tasks and the competency construct. Therefore, the evidence model consists of two sub-models, that is (1) the evidence rules and (2) the statistical model (i.e. computational model). The evidence rules describe the relationship between the observed performances and the in-game tasks, while the statistical model describes the relationship between the observed performances and the competency construct.

Within the scope of this study, the only relevant models are the competency model and the statistical model, since only these two models are essential for the evaluation of the learners' performance from logged data. The task model and the evidence rules become important only when the SA is to be integrated within the game source code itself, which requires close attuning of the game's design to these models. GSAT, however, does not concern about these aspects as it exclusively deals with the diagnostic aspect of SA and its generic application even in games that have not been developed with respect to ECD.

2.2 Machine Learning Technology

Serious games are frequently portrayed as one of the most promising digital vehicles for capturing rich learner data, far beyond of what is usually possible in traditional education settings. This rich data can be used to fathom the behaviour of the learners and evaluate their competence level even for imponderables such as soft skills (i.e.

communication, collaboration, team-work, etc.) and even personality traits. Machine learning is a field artificial intelligence that uses data to build models for pattern recognition and inferences. For SA, ML is the most suitable technology for making predictions about competence levels from logged data. Originally, Bayesian Networks were examined as an ML methodology for SA [2]; however other ML algorithms such as Decision Trees, Neural Networks, Logistic Regression, Support Vector Machines, and Deep Learning have been explored for SA [14, 15].

3 GSAT

The main motivation for developing GSAT was to lift the barriers of SA and allow its wider application in serious games. While SA served its purpose well in several case-specific empirical studies, the concept of directly integrating it within the gaming environment has hindered its full potential. Hence, the idea of detaching SA from hardcoded solutions led to developing GSAT as a practicable stand-alone software tool. Fundamentally that was possible due to the generic nature of the main ingredients that constitute SA (i.e. ECD and ML). As a result, GSAT not only allows the wider adoption of SA by the serious game community, but also offers research opportunities for examining SA when exposed to various boundary conditions.

GSAT was developed as a client-side console application in the C# programming language using the .NET framework. Currently an early version has been developed which fulfils all its core functional requirements [9], be it without a user interface, help widgets, and additional support functions (future work will address these issues to enhance the usability of the tool). GSAT's workflow design, as well as the external libraries that were used to realize it, has been extensively presented in a previous study [10], which used a simulation-based approach to examine the robustness of GSAT for numerical datasets of different sample sizes and normality significance levels, for different competency constructs and statistical models, and when using different ML algorithms. The results have shown that GSAT is a highly robust tool that ranked high in all the used performance measures for all the tested conditions.

4 Big Five Personality Traits

Since early 20th century, efforts were made concerning the development of a descriptive model for personality. These efforts led to a five factor model [16], referring to the following factors: (1) Openness to new experiences, (2) Conscientiousness, (3) Extraversion, (4) Agreeableness, and (5) Neuroticism (abbreviated to OCEAN). Accordingly, a valid and reliable test instrument for the OCEAN personality traits is available: the NEO PI-R questionnaire. The NEO PI-R divides every trait into six facets (see Table 1) and consists of 240 items measuring the five domains and their facets. In the reference study [11] that provided us with the *THE POISONED LAKE* datasets, data was also collected using the NEO PI-R questionnaire. The final scores of the learners that participated in this study were normed according to a respective valid norm table that takes into account the distributions on large sample groups.

Table 1. The five personality traits followed by their general description and respective facets.

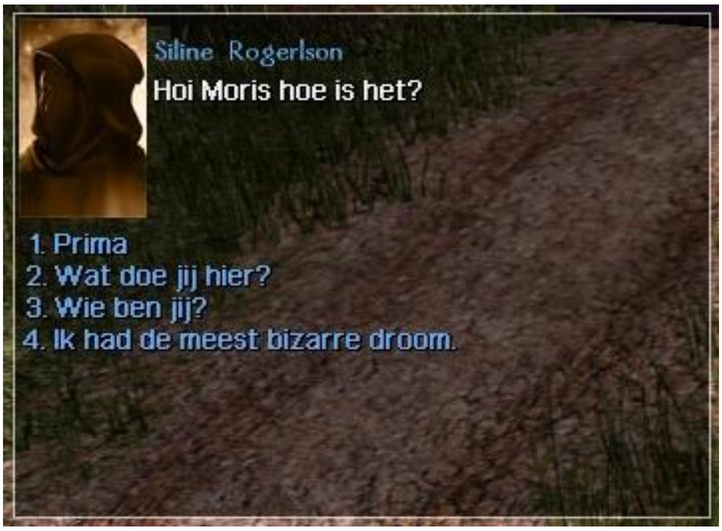
Personality Traits	Description	Facets
Openness	The interest in novel stimuli. A high score is typically accompanied by curiosity and willingness to deviate from social conventions.	Fantasy Aesthetics Feelings Actions Ideas Values
Conscientiousness	The propensity to adhere to rules, both social and personal. This trait is also tied to the ability to restrain oneself and the ability to stick to a plan during periods of stress and difficulty.	Competence Order Dutifulness Achievement Striving Self-Discipline Deliberation
Extraversion	High scorers seek excitement and positive stimuli. This often leads to individuals seeking the company of others and seeking exhilarating situations like high speed driving, roller coasters, and other high adrenaline activities.	Warmth Gregariousness Assertiveness Activity Excitement Seeking Positive Emotion
Agreeableness	Explained as compliance, willingness to cooperate, and friendliness. Low scorers tend to follow their own needs over those of others. High scorers are seen as empathic.	Trust Straightforwardness Altruism Compliance Modesty Tendermindedness
Neuroticism	This trait is connected to fluctuating and negative emotions such as anger and fear (see Figure 2.1). High scorers are more likely to check situations for safety. There is also a relationship to shyness and social anxiety.	Anxiety Hostility Depression Self-consciousness Impulsiveness Vulnerability to Stress

5 THE POISONED LAKE

THE POISONED LAKE game (see Fig. 1) was developed as a mod of the popular leisure game called *NEVERWINTER'S NIGHT*. Information regarding the gameplay, the data that was logged during gameplay, and the statistical models that were finally produced from this data as reported in [11] are presented below.

5.1 Gameplay

The gameplay of *THE POISONED LAKE* involves a storyline that is divided into three parts: (1) a training part so that the learners become familiar with the game controls, (2) the main part at which the learners have to execute a mission of solving the mystery of the poisoned lake and finding a way to stop the poisoning, (3) a couple of optional side stories were the learners can investigate how to save various non-playing characters (NPCs). The main actions that learners can perform during gameplay are to venture in the map and converse with NPCs in order to find a way to solve the mystery. The maximum amount of gameplay time for the learners was set to 60 minutes.



Siline Rogerlson

Hoi Moris hoe is het?

1. Prima
2. Wat doe jij hier?
3. Wie ben jij?
4. Ik had de meest bizarre droom.

Fig. 1. A screenshot from *THE POISONED LAKE* game were a learner talks to an NPC [11].

5.2 Game Logs

Discrete numerical data was logged during gameplay for 80 learners (same for the NEO PI-R questionnaire). The logged data referred to three distinct types of variables: (1) data related to conversations with NPCs, (2) data related to the movement of the learners in the map logged at certain trigger points, and (3) general data relating the total time spend in game as well as aggregated (i.e. pooled) data regarding both conversation and movement data. A total of 260 game variables were logged [11].

5.3 Statistical Models

A linear stepwise regression analysis was performed in order to generate statistical models for each personality based on the collected game data. To achieve this, the final normed scores of the learners for each of the big five personality were set as dependent variables, while all the logged game variables were set as independent variables. Hence, five statistical models were generated each one explaining a certain amount of the variance by the models according to the size effect statistic R^2. Table 2 depicts the results of the stepwise regression analysis including validity and reliability relevant statistics (R^2 and Cronbach's α) [11].

Table 2. Overview of the statistical models produces from the linear stepwise regression analysis based on the logged for the big five personality traits [11].

Personality Traits	R^2	α	No. of variables in model
Openness	.768	.54	17
Conscientiousness	.559	.31	10
Extraversion	.351	.07	6
Agreeableness	.724	.07	15
Neuroticism	.568	.55	9

6 Methodology

6.1 Using Statistical Models for the Big Five Personality Traits with GSAT

We configured GSAT to run the statistical models that were provided by the reference study in order to produce inferences about the personality traits of the learners. A Gaussian Naïve Bayesian Network (GNBN) was used for each personality trait. A percentage split rule was used to decide the number of samples included for the training (65%) and testing (35%) purposes of the classifiers. The GNBNs were set to produce inferences for three classes (Low, Medium, and High performance). Several performance measures [17] were used in this study to evaluate the performance of GSAT, such as the classification accuracy (CA), the kappa statistic (KS), the mean absolute error (MAE), the root mean squared error (RMSE), the relative absolute error (RAE), and the root relative squared error (RRSE).

6.2 Validation of the Results

A bivariate correlation analysis approach is used in this study in an attempt to validate the outcomes of GSAT regarding the big five personality traits of the learners. That is, we examined the Spearman's *rho* correlation coefficients between the normed results from the NEO PI-R questionnaire and the classifications produced by GSAT.

7 Results

This section includes results relating to both the performance of GSAT and the validity of the used statistical models.

7.1 GSAT performance

The results for each GNBN classifier used per personality trait can be found in Table 3.

Table 3. Results regarding the performance of GSAT on the big five personality traits.

Personality Traits	CA	KS	MAE	RMSE	RAE (%)	RRSE (%)
Openness	0.96	0.94	0.04	0.19	6.1	26.7
Conscientiousness	0.74	0.51	0.26	0.51	56.3	85.3
Extraversion	1	1	0	0	0	0
Agreeableness	1	1	0	0	0	0
Neuroticism	0.78	0.64	0.22	0.47	35.16	64.0

7.2 Correlation analysis results

A bivariate correlation analysis between the outcomes of GSAT and NEO PI-R was performed for each of the big five personality trait of the OCEAN model in order to validate the GSAT's outcomes with respect to the used statistical models. The results of this analysis are depicted in Table 4.

Table 4. Results from the bivariate correlation between the outcomes from NEO PI-R and GSAT with respect to Spearman's *rho* coefficient. The ** sign suggests significant correlation at the 0.01 level (2-tailed).

Personality Traits	Spearman's rho
Openness	-.104
Conscientiousness	-.099
Extraversion	-.270
Agreeableness	.504**
Neuroticism	.357

8 Discussion and conclusion

This study examined GSAT's performance with real-world data collected from a serious game. When examining the performance of GSAT by using standard classification performance measures it was found that the GNBN classifiers were able to perform at a high level despite the small sample size. Most notably, high classification accuracies (100%) were found for extraversion and agreeableness, while the lowest classification accuracy was found for conscientiousness (74%). These results confirm the robustness of GSAT with real-world data.

The bivariate correlation analysis of the GSAT outcomes with the respective outcomes from the NEO PI-R shows a strong and significant correlation only for agreeableness. This is reassuring as such, be it only a partial success. The reason for not being able to validate all the statistical models may lie on possible overfitting issues in the original model. Another possible explanation is that the original statistical models were not fully explaining the variance dependent variables in the first place. Indeed, an additional analysis of the data has revealed some flaws. In specific, we examined regression assumptions such as linearity, collinearity, normality, outliers, etc. Not to mention that the sample size was probably too small [18, 19, 20] for the number of descriptors included in the regression.

Nevertheless, this study was an excellent opportunity to test GSAT with real-world data. Even with a small dataset (only 80 users) GSAT was capable of training the SA model, and making inferences on the users' personality traits, be it only partially. This provides a favourable starting point for follow-up studies with larger sample sizes and more reliable statistical models. Moreover, by testing GSAT this study as demonstrated the practicable feasibility of the SA methodology. It also has shown that the generation of valid and reliable statistical models is essential for full and reliable coverage of assessments in serious games. Overall, this study contributes to improving the visibility, feasibility, and practicability of a principled assessment methodology for serious games such as SA.

References

1. Larson, L. C., Miller, T. N.: 21st century skills: Prepare students for the future. Kappa Delta Pi Record, 47(3), 121-123 (2011).
2. Shute V. J.: Stealth assessment in computer-based games to support learning. Computer games and instruction 55.2: 503-524 (2011).
3. Mislevy. R. J.: Evidence-Centered Design for Simulation-Based Assessment. CRESST Report 800. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (2011).
4. Shute, V. J., Ventura, M., Kim, Y. J.: Assessment and learning of qualitative physics in newton's playground. The Journal of Educational Research 106.6: 423-430 (2013).
5. Ventura, M., Shute, V., Small, M.: Assessing persistence in educational games. Design recommendations for adaptive intelligent tutoring systems: Learner modeling 2: 93-101 (2014).
6. Shute, V. J., Wang, L., Greiff, S., Zhao, W., Moore, G.: Measuring problem solving skills via stealth assessment in an engaging video game. Computers in Human Behavior, 63, 106-117 (2016).
7. Moore, G. R., Shute, V. J.: Improving learning through stealth assessment of conscientiousness. In Handbook on digital learning for K-12 schools (pp. 355-368). Springer, Cham. (2017).
8. Georgiadis, K., Van Lankveld, G., Bahreini, K., & Westera, W..: Accommodating Stealth Assessment in Serious Games: Towards Developing a Generic Tool. In 2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games) (pp. 1-4). IEEE. (2018).
9. Georgiadis, K., Van Lankveld, G., Bahreini, K., & Westera, W..: Learning Analytics Should Analyse the Learning: Proposing a Generic Stealth Assessment Tool. Accepted at the IEEE Conference on Games (CoG). (2019).

10. Georgiadis, K., Van Lankveld, G., Bahreini, K., & Westera, W..: On The Robustness of Steath Assessment. Submitted to IEEE Transactions on Games. (2019)
11. Van Lankveld, G., Spronck, P., Van den Herik, J., Arntz, A.: Games as personality profiling tools. In 2011 IEEE Conference on Computational Intelligence and Games (CIG'11) (pp. 197-202). IEEE. (2011).
12. McCrae, R. R., Costa Jr, P. T.: Personality trait structure as a human universal. American psychologist, 52(5), 509. (1997).
13. Costa, P. T., McCrae, R. R.: The revised neo personality inventory (neo-pi-r). The SAGE handbook of personality theory and assessment, 2(2), 179-198. (2008).
14. Sabourin, J. L.: Stealth Assessment of Self-Regulated Learning in Game-Based Learning Environments. (2013).
15. Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., Lester, J. C.: DeepStealth: leveraging deep learning models for stealth assessment in game-based learning environments. In International Conference on Artificial Intelligence in Education (pp. 277-286). Springer, Cham. (2015).
16. Wiggins, J. S. (Ed.): The five-factor model of personality: Theoretical perspectives. Guilford Press. (1996).
17. Domingos, P. M.: A few useful things to know about machine learning. Commun. acm, 55(10), 78-87. (2012).
18. Field, A.: Discovering statistics using SPSS. Sage publications. (2009).
19. Green, S. B.: How many subjects does it take to do a regression analysis. Multivariate behavioral research, 26(3), 499-510. (1991).
20. Maxwell, S. E.: Sample size and multiple regression analysis. Psychological methods, 5(4), 434. (2000).