Bolstering Stealth Assessment in Serious Games

Konstantinos Georgiadis[1], Tjitske Faber[2], Wim Westera[1]

[1]*Open University of the Netherlands, 6419AT Heerlen, The Netherlands, {konstantinos.georgiadis, wim.westera}@ou.nl*

[2] *Erasmus University of the Netherlands, 3015GD Rotterdam, The Netherlands, t.faber@erasmusmc.nl*

Abstract

Stealth assessment is an unobtrusive assessment methodology in se-rious games that use digital player traces to make inferences of players' exper-tise level over competencies. Although various proofs of stealth assessment's validity have been reported, its application is a complex, laborious, and time-consuming process. To bolster the applicability of stealth assessment in serious games; a generic stealth assessment tool (GSAT) has been proposed, which us-es machine learning techniques to reason over competence constructs, player log data and assess player performance. The current study provides empirical validation of GSAT by applying it to a real-world game, the abcdeSIM game, which was designed to train medical care workers to act effectively in medical emergency situations. GSAT demonstrated, while relying on a Gaussian Naive Bayes Network, to be highly robust and reliable, achieving a three-level as-sessment accuracy of 96.8 %, as compared with a reference score model defined by experts. By this result, this study contributes to the alleviation of stealth assessment's applicability issues and hence promotes its wider uptake by the serious game community.

Keywords: Stealth Assessment, Generic Tool, Statistical Model, Machine Learning, Stepwise Regression, Serious Games, ABCDE-method.

1 Introduction

As opposed to traditional classroom teaching, the use of serious games yields a bulk of digital traces of learner actions, which open up new opportunities for the in-depth assessment of learners: so-called stealth assessments (SA), unobtrusively based on the learner's behavioural traces in the game, without the need for explicit test items.

SA is grounded in a principled methodology [1], combining a design framework for modeling the assessment process and advanced data science technologies. Evi-dence-Centered Design (ECD) [2] has served as the design framework that provides a generic layout for developing competency, task, and evidence (i.e. data) constructs. Most notably, these constructs lead to

the development of statistical models describing the relationships between competencies, tasks, and evidence, which can be used to meaningfully represent gameplay data within machine learning (ML) algorithms to produce valid inferences (i.e. classifications). Several empirical studies [3, 4, 5] have provided a proof of concept for SA.

Notwithstanding the proof that SA can indeed provide valid and reliable assess-ments, its practical application turns out to be complex, laborious, and time-consuming. Applying SA in serious games requires substantial expertise in different domains including assessment, data science, statistics, game development, machine learning, etc. Until recently, no tools existed to automate the data processing part of SA in a generic way, thus forcing anyone who wants to apply SA to manually develop it from scratch in a hardcoded manner. This not only turns applying SA into a time-consuming process, but it also reduces its transferability to other serious games, in-creases development costs, and makes it prone to mistakes. Overall, these practical barriers turn an otherwise exquisite assessment methodology into an unattractive as-sessment alternative for a wide spectrum of the serious game community.

In a recent study, a generic tool for applying SA has been proposed [6, 7], which tackles these issues and allows its broader application in serious games. The Generic Stealth Assessment Tool (GSAT) is a stand-alone software application that (a) handles numerical datasets from any serious game, (b) automatically runs ML processes, and (c) allows the easy arrangement of diverse ECD models. GSAT has been successfully tested for its robustness against several conditions with simulation datasets of different sizes and normality significance levels, with different ML algo-rithms, and different ECD models [8]. The current study goes beyond simulation data: it examines the robustness and reliability of GSAT with real-world data collected from the serious game abcdeSIM. This game supports healthcare practitioners and students at the acquisition of emergency medical care skills. The players are trained in properly applying an emergency treatment method called the ABCDE-method. A detailed rating system developed by experts was implemented within the game to evaluate the players' performances by registering points to them after every correct or false action. Using these expert scores, a benchmark for assessment allowed us to study and compare the quality and validity of the SA approach that was established using GSAT.

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 in-formation on the ABCDE-method can be found in section 4. Details regarding the game and the collected data 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 mentioned above, SA is essentially the combination of a principled conceptual framework for designing data-driven assessments in serious games. It relies on the ECD framework along with ML algorithms.

2.1 Evidence-Centered Design

The ECD [2] is a framework for designing assessments in serious games on the basis of several generic conceptual models. ECD describes the assessment design as part of three such models. These are: (a) the competency model, which describes the assessed competency and its underlying facets (i.e. sub-skills), (b) the task model, which describes the in-game tasks that can elicit data relative to the assessed competency con-structs, and (c) the evidence model, which describes the relationships of the elicited data (i.e. game variables / observables) to both the tasks in the game and the compe-tency constructs. In particular, the relationship between game data and competency constructs, which is commonly referred to as the statistical model, is crucial for apply-ing SA. GSAT was designed to solely focus on the assessment aspects of ECD (such as the statistical and competency models).

2.2 Machine Learning

ML, being a subset of artificial intelligence technology, is considered to be on the verge of data science due to its capability for providing probabilistic solutions to non-binary and non-deterministic problems. Generally speaking, a wide array of ML al-gorithms exists that are usually categorized as supervised, unsupervised, or semi-supervised learning algorithms. Supervised ML algorithms produce inferences by using as reference point a set of labeled training data (e.g. a pre-annotated dataset with classifications by experts). Unsupervised ML algorithms can provide classifica-tions by applying clustering techniques on unlabeled datasets. Semi-supervised ML algorithms fall in-between the previous two categories, where both labeled and unla-beled data are used to provide classifications. The ML algorithms belonging to these categories can be further distinguished according to the different types of data (e.g. numerical, categorical, ordinal, etc.) that they can process.

As mentioned previously, serious games offer the opportunity to capture vast amounts of data far beyond a teacher's grasp within a traditional classroom. After structuring this data into meaningful statistical models (e.g. through ECD) allows for the use of ML algorithms to assess (i.e. classify) the learners' performances. So far, various ML algorithms (such as Bayesian Networks, Support Vector Machines, Deci-sion Trees, and Deep Learning) have been used for SA [9, 10]. Nevertheless, Bayesi-an Networks have been the first and foremost used algorithm in existing SA studies [3, 4, 5] probably due to its generative nature. In this study, we opted for a

Gaussian Naive Bayes Network (GNBN) as we deal with numerical data in a supervised man-ner. Also, GNBN has turned out to be the highest performing algorithm in our simulation studies with GSAT [8].

3 GSAT

SA is essentially a generic methodology as its two components (i.e. ECD and ML) are generic in nature by default. However, so far SA has only been applied in a hardcoded manner, that is, directly implemented within a game's source code and thus specifically destined for assessing a certain competency within the context of a specific game. GSAT offers a generic tool to define and operate SA across a diversity of com-petencies and game contexts. GSAT is currently a stand-alone software solution that offers fully automated processing of numerical datasets with ML, and that exclusively deals with the diagnostic aspects of SA. It allows for declaring the competency model and once it is fed with player log data it uses ML classifiers to deliver the player-related performance assessments. The tool is a client-side console application devel-oped in the C# programming language using the .NET framework. A more detailed description of GSAT including its technical architecture, workflow design, external libraries, etc. can be found in previous studies [6, 7].

4 The ABCDE-method

The ABCDE-method is an emergency medical care protocol used by healthcare providers all over the world for assessing and treating acutely ill patients. This meth-od relies on the principle of "treat first what kills first". This means that the healthcare practitioner needs to follow a systematic approach when dealing with acutely ill pa-tients. This approach is divided into five phases, using a simple mnemonic: Airway, Breathing, Circulation, Disability, and Exposure/Environment (ABCDE). See Table 1 for a brief description of each phase.

Table 1. Descriptions of the five phases of the ABCDE-method (not exhaustive).

Phases	**Description**
Airway	Check for abnormalities indicating airway obstruction by addressing the patient, listening for abnormal breathing and inspecting the oral cavity for blood, vomit, swelling. In case of obstruction, manual airway manoeuvers or airway devices should be used. If the airway is blocked, specialist help should be called.

Breathing	Check for pulmonary disorders (e.g. pneumonia, asthma, etc.). If ventilation or oxygenation is inadequate, put the patient in a sitting position, administer oxy-gen and treat the underlying cause. Consider manual or mechanical ventilation in severe cases.
Circulation	Check for loss of circulating volume, decreased cardiac contractility, and loss of vascular pressure. In case of disorders, establish intravenous access, and, depending on the underlying issue, administer fluids or vasoactive medication. Specific cases require consultancy from specialists.
Disability	Evaluate neurological status by examining consciousness, meninges, pupillary response, and glucose levels. Look for signs of intracranial hemorrhage, intoxi-cation, hypoglycemia, or electrolyte disorders. Treat the underlying disorder, with specialist help if needed. Protect the airway: obstruction may occur when consciousness deteriorates.
Exposure / Environment	Look for skin and other physical abnormalities (e.g. injury signs, rashes). Treat disorders that require immediate attention, consult appropriate specialists, and cover the patient to avoid hypothermia.

5 The Game: abcdeSIM

The ABCDE-method is generally taught to healthcare practitioners in face-to-face courses across several contexts (e.g. trauma, medicine, obstetrics, and pediatric cours-es). Although such courses are generally effective, still there is room for improvement since the costs are high [11] and the opportunities for distributed practice are limited [12]. It has been suggested to use a serious game for addressing these issues, which would also allow teaching complex cognitive skills in an engaging, flexible and pa-tient-safe way [13].

Therefore, a serious game called abcdeSIM was developed in close collaboration between medical practitioners, game designers, and educationalists from the Erasmus University Medical Center, Rotterdam, and VirtualMedSchool, Rotterdam. The aim of the game is to prepare residents in emergency medicine care by applying the ABCDE-method. The abcdeSIM game has already been used in research [14] and training. Now, the game is used for testing and validating the SA produced by GSAT.

5.1 Gameplay

In the abcdeSIM game, the player takes on the role of a physician who is presented with an acutely ill patient in a virtual emergency department. The virtual nurse pro-vides a brief

handover containing information on patient's condition. All tools and information available in a real-life emergency department are available to the player. Among other things, the player can perform physical examinations, talk to the patient, administer medication, order diagnostic tests, and ask for help from a specialist. In fifteen minutes, the player must complete a full examination of the patient and initiate necessary treatments.

Fig. 1. A snapshot from the abcdeSIM gameplay.

Vital parameters and the condition of the patient are generated by a complex phys-iological model that is influenced by the player's actions. This results in a realistic feel of the scenario (Fig. 1). Afterward, the player can choose to proceed on a "secondary survey". At this point, corrective feedback, a game score based on an expert rating system, and a narrative on how the patient fared after their care are provided. Several patient cases are available with different medical conditions and levels of sickness.

For the purpose of this study, a game scenario concerning a patient suffering from subarachnoid hemorrhage was used. The patient presents with an obstructed airway, necessitating the use of the ABCDE-method to improve her condition swiftly. Before playing this scenario, players are advised to first follow a gameplay tutorial, complete a practice scenario without any illness, and apply their skills in an emergency scenar-io. This ensures their familiarity with the game interface.

5.2 Game Logs

For this study, we looked at log files of first attempts at completing the scenario containing anonymized raw data collected during gameplay from 267 players. Each log file was parsed using a specialized JavaScript parser, which allowed extracting and categorizing logged events based on actor (player or patient), action type (exami-nation, exploration, intervention, history, diagnostics, help-seeking from specialist, reflection), and the concurrent ABCDE-method phase if applicable. From this parsed log we calculated the total number of actions for the player and patient and the total number of actions per action type. In addition, aggregated game variables were devised for each action type, that is, the ratio of the number of times each action type was performed to the total number of player's actions. Finally, to quantify adherence to the ABCDE-method, we calculated systematicity scores for each session using a Hidden Markov Model as described by Lee and colleagues [15].

The rating system (developed by content experts) embedded within abcdeSIM allowed for logging final game scores for each player. The game score depends on the number of correct or false decisions made according to the ABCDE-method. Harmful interventions (e.g. administering the wrong medication) subtract points, while helpful interventions add points to the players' final game score. Completing a case faster than fifteen minutes rewards additional points.

6 Methodology

6.1 Statistical Model

For deriving a meaningful statistical model from the logged data it is important to first describe a competency construct that specifies the underlying abilities needed to properly apply the ABCDE-method. During each of the ABCDE-method's phases, the players have to perform a series of actions that require medical and game procedural abilities. The medical part covers three theoretical aspects of the medical care process: (a) ability to properly diagnose the medical problem, (b) ability to apply appropriate treatments, and (c) ability to systematically follow the ABCDE-method and reflect on the outcomes of each phase. The game procedural part relates to the practical aspects of performing the correct actions (e.g. navigating the game environment, selecting tools, applying the tools to the correct areas).

To investigate and establish the statistical relationships between the game variables and the competency construct a linear stepwise regression analysis on the data was applied. Beforehand, certain regression assumptions were first examined, such as the linearity of the relationship of the game variables (independent variables) and the final game scores from the expert rating system (dependent variable), as well as the collin-earity between the game

variables. In this way, we minimized the set of game varia-bles that should be included in the regression analysis and thus maintain only the most influential ones. Finally, five game variables made it into the regression analysis: (1) relative examination ratio: no. of examinations to no. of player actions, (2) relative reflection ratio: no. of reflections to no. of player actions, (3) relative diagnostics ratio: no. of diagnostics to no. of player actions, (4) systematicity, and (5) inversed relative exploration ratio: inversed no. of exploration to no. of player actions. The model turns out to explain 53% of the variance (Adjusted R2=0.53) of the dependent variable and shows an acceptable level of internal consistency (Cronbach's α=0.602). Fig. 2 provides a view of the statistical model.

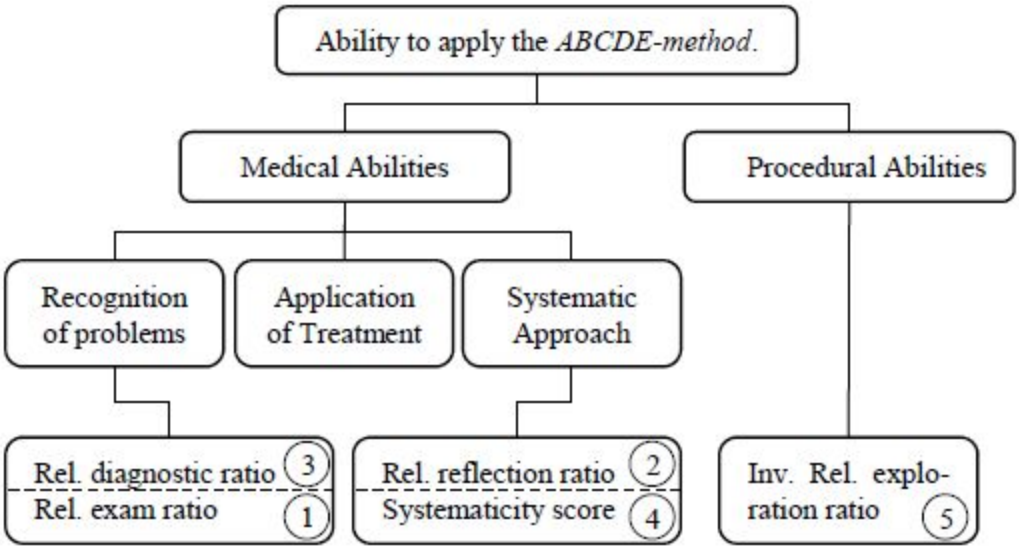

Fig. 2. A view of the statistical model for the players' ability to apply the ABCDE-method.

6.2 GSAT's Configuration and ML Performance Measures

In order to classify the player's ability to apply the ABCDE-method, we imported the statistical model to GSAT along with the respective data. Since the data originally was not labeled (by experts or otherwise), a clustering approach described in [8] was first applied to label the data. Then, a GNBN algorithm was used to produce infer-ences with regard to three classes (Low, Medium, and High performance). A percent-age split rule was used to train (65% of the samples used) and test (remaining 35% of the samples used) the classifier. In accordance with to [16] several performance measures were used to evaluate the performance of the GNBN classifier including 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), respectively.

6.3 Validation Process

To validate the outcomes of GSAT we examined the Spearman's rho correlation coef-ficients of the classifications produced by GSAT and the game scores from the expert rating system. To this end, the game scores were first clustered using a k-means clus-tering approach (3 clusters) in order to have both variables aligned in an ordinal form.

7 Results

7.1 GSAT's Performance

The various performance measures (cf. section 6.2) for the GNBN classifier are presented in Table 2.

Table 2. GSAT's robustness according to GNBN's performance measures.

ML	CA (%)	*KS*	*MAE*	*RMSE*	*RAE (%)*	RRSE (%)
GNBN	96.8	0.94	0.03	0.18	4.11	21.01

7.2 Validation of GSAT's Outputs

A bivariate correlation analysis between the game scores from the expert rating sys-tem and the classifications produced for the players' ability to apply the ABCDE-method was performed in order to validate GSAT's outcomes. The result of this anal-ysis suggests a significant correlation of the two given by the Spearman's rho coeffi-cient of 0.607 at a p=0.01 significance level (2-tailed).

8 Discussion and Conclusions

In this study we managed to derive a statistical model effectively describing mean-ingful relationships between the collected log data and the abilities that are needed to apply this ABCDE-method in the abcdeSIM game environment. The statistical model displayed internal consistency at an acceptable level (Cronbach's α=0.602). Entering this model into GSAT allowed for making inferences regarding the players' perfor-mances. The classifications produced by GSAT are highly correlated (Spearman's rho = 0.607 at a 0.01 significance level) with the expert scores. Concerning GSAT's performance, we found that when applying a GNBN

algorithm on the declared statistical model its classification accuracy (being the most important performance measure) is 96.8%. This means that the classifier was able to accurately assess most of the tested cases.

A substantial limitation of the study is the limited size of the dataset. As a conse-quence, the statistical power and accordingly the number of behavioral predictors (cf. section 6.1) are constrained. Notably, in this setting, no relevant game variable could be identified that would cover applying proper treatments. One explanation is that a portion of data (e.g. certain mouse clicks relating to dosage or response to double-check procedure) that could potentially relate to this facet was not logged in the first place. Another explanation could be that after correctly diagnosing the patient condi-tion, players naturally follow the indicated treatment procedures so a strong covari-ance with examination actions exists. A third explanation is that the abcdeSIM game does not teach detailed medical procedures (such as placing an i.v. cannula or chest tube), but is most effective in practicing the cognitive skill of following the appropri-ate steps of the ABCDE method, with limited operational details. Overall, abcdeSIM offered an excellent opportunity to detail the practical application of the SA method with GSAT, demonstrating the potential of automation tools that can lift the barriers of SA and allow its wider application in the domain of serious games.

Acknowledgments

We wish to thank IJsfontein, a serious game company in Amsterdam, for making the game and extensive logging utility available and Virtual MedSchool for providing (anonymized) game log data. We also acknowledge Jeroen Donkers of the School of Health Professions Education, Maastricht University, the Netherlands, for assisting at calculating systematicity scores and data processing. Finally, we thank Tin de Zeeuw, Lent, the Netherlands, for his assistance in processing the raw game log data.

References

1. Shute V. J.: Stealth assessment in computer-based games to support learning. Computer games and instruction 55.2: 503-524 (2011).

2. 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).

3. 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).

4. Ventura, M., Shute, V., Small, M.: Assessing persistence in educational games. Design recommendations for adaptive intelligent tutoring systems: Learner modeling 2: 93-101 (2014).

5. 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).

6. 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).

7. 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).

8. Georgiadis, K., Van Lankveld, G., Bahreini, K., & Westera, W..: On The Robustness of Stealth Assessment. Submitted to IEEE Transactions on Games. (2019)

9. Sabourin, J. L.: Stealth Assessment of Self-Regulated Learning in Game-Based Learning Environments. (2013).

10. 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).

11. Perkins, G. D., Kimani, P. K., Bullock, I., Clutton-Brock, T., Davies, R. P., Gale, M., ... & Stallard, N.: Improving the efficiency of advanced life support training: a randomized, controlled trial. Annals of internal medicine, 157(1), 19-28. (2012).

12. Cook, D.A.; Hamstra, S.J.; Brydges, R.; Zendejas, B.; Szostek, J.H.; Wang, A.T.; Erwin, P.J.; Hatala R (2012) Comparative effectiveness of instructional design features in simula-tion-based education: systematic review and meta-analysis. Med Teach 35:e867-898.

13. Kalkman, C. J.: Serious play in the virtual world: can we use games to train young doc-tors?. Journal of Graduate Medical Education, 4(1), 11-13. (2012).

14. Dankbaar, M. E., Roozeboom, M. B., Oprins, E. A., Rutten, F., van Merrienboer, J. J., van Saase, J. L., & Schuit, S. C.: Preparing residents effectively in emergency skills training with a serious game. Simulation in Healthcare, 12(1), 9. (2017).

15. Lee JY, Donkers J, Jarodzka H, van Merriënboer JJG (2019) How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. Comput Human Behav 99:268–277.

16. Domingos, P. M.: A few useful things to know about machine learning. Commun. acm, 55(10), 78-87. (2012).