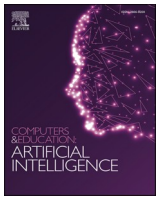
Computers and Education: Artificial Intelligence 8 (2025) 100351

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Adaptive serious games assessment: The case of the blood transfusion game in nursing education

Dirk Ifenthaler [a,b,*], Muhittin Sahin [b], Ivan Boo [c], Darshini Devi Rajasegeran [d], Ang Shin Yuh [d]

[a] Curtin University, Perth, Australia
[b] University of Mannheim, Mannheim, Germany
[c] Serious Games Asia, Singapore
[d] Singapore General Hospital, Singapore

1. Introduction

Serious games have gained prominence in education and training (Loh et al., 2015; Squire, 2006), with research increasingly exploring their potential for learning (Laamarti et al., 2014). A serious game is broadly defined as a digital game created not with the primary purpose of pure entertainment but with the intention of serious use in training and education (Loh et al., 2015; Squire, 2006). Learning success within serious games has been linked to factors such as backstory and production, realism, interaction, feedback and debriefing, as well as the integration of artificial intelligence and adaptivity into the serious game (Ravyse et al., 2017). Only recently, artificial intelligence (AI) has been linked with serious games (Tolks et al., 2024). For instance, AI approaches enable advanced game-play architecture and the creation of Non-Playable Characters (NPCs) with more sophisticated behaviours and decision-making abilities (González-Calero & Gómez-Martín, 2011).

While factors like adaptivity and artificial intelligence integration enhance game-based learning (Aydin et al., 2023; Ravyse et al., 2017), the use of serious games for assessment remains limited (Ge et al., 2021; Kim & Ifenthaler, 2019; Smith et al., 2015).

Unlike traditional assessments, serious games generate rich gameplay data that can inform real-time performance analysis (Loh & Sheng, 2014). AI can further enhance this potential by enabling advanced game mechanics and personalised feedback (González-Calero & Gómez-Martín, 2011; Hildmann, 2024).

Despite successful implementations in various fields, for instance, programming, physics, mathematics, computational thinking, logic, and health (Boyle et al., 2016), the application of serious games assessment in nursing education and their workplace is nascent (Sánchez-Valdeón et al., 2023; Sánchez-Valdeón et al., 2023; Thangavelu et al., 2022; Thangavelu et al., 2022).

Hence, this study will explore and improve the understanding of adaptive serious game assessment approaches in nursing education at the workplace. More precisely, this study investigates whether gameplay data accurately reflects player interaction and whether an adaptive assessment algorithm can effectively identify different competence levels during gameplay, comparing its results to a traditional game score.

1.1. Serious games assessment

Serious games were first formally defined in Clark Abt's 1970s book "Serious Games" (Abt, 1970). The term originated from the author's game development, focusing on simulation games for skills acquisition. Since then, there have been several attempts to define *serious games*. For instance, Abt (1987, p. 9) broadly suggests that serious games "have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement". Zyda (2005, p. 26) defines serious games linked to use cases as "mental contests played with a computer in accordance with specific rules that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives". The definition by Ritterfeld et al. (2009, p. 3) addresses the multiple objectives of serious games "as involving fun, as well as being educational, engaging, impactful, meaningful, and purposeful". Hence, serious games are a still-emerging field where (mostly) digital games are supposed to use sound learning theories and learning design principles to support learning processes and foster learning outcomes. In a widely cited meta-analysis, Clark et al. (2016) report that serious games had a moderate to strong effect on improving overall learning outcomes, including cognitive and interpersonal skills, compared to non-game conditions.

Serious games offer a dynamic learning environment to assess various skills, knowledge, and competencies more engagingly and

* Corresponding author. University of Mannheim, Mannheim, Germany
E-mail addresses: dirk@ifenthaler.info (D. Ifenthaler), muhittin.sahin@uni-mannheim.de (M. Sahin), ivan@seriousgamesasia.com (I. Boo), darshini.devi.rajasegeran@sgh.com.sg (D.D. Rajasegeran), ang.shin.yuh@singhealth.com.sg (A.S. Yuh).

https://doi.org/10.1016/j.caeai.2024.100351
Received 26 March 2024; Received in revised form 16 December 2024; Accepted 21 December 2024
Available online 26 December 2024

interactively compared with traditional assessment methods (Caballero-Hernández et al., 2017). Serious games are being implemented to assess cognitive abilities and problem-solving skills (Sánchez & Olivares, 2011), decision-making processes (Loh et al., 2016), emotion regulation (Jerčić & Sundstedt, 2019), or teamwork (Vlachopoulos & Makri, 2017). These games often provide real-time feedback, allowing educators to gauge a student's progress more effectively. Further, a significant benefit of serious games assessment is the capacity to interact with real-world scenarios. This enables learners to apply their knowledge and skills in practical situations, providing a more authentic assessment. For instance, serious games are used in the health sciences to assess a learner's ability to handle complex medical situations (Kitchen & Humphreys, 2014; Tsopra et al., 2020). Moreover, serious games allow assessments to be tailored to specific learning objectives or competencies (Ge & Ifenthaler, 2017).

However, when it comes to serious games assessment, Ifenthaler et al. (2012) argued that implementing assessment features into serious games is slowly progressing because it adds a very time-consuming step to the design process of serious games. Several distinguishing features of assessment with serious games have been proposed and are widely accepted: (a) game scoring, (b) external assessment, and (c) embedded assessment with serious games (Ifenthaler et al., 2012). Caballero-Hernández et al. (2017) confirm that game scoring and embedded assessment combined with external assessment have been most frequently used in serious games. An additional feature includes analytics of game data, specifically denoted as serious games analytics (Loh et al., 2015). Serious games analytics converts learner-generated data into actionable insights for real-time processing and decision-making. Metrics for serious games analytics include learner-generated game data (e.g., time spent, obstacles managed, goals or tasks completed, navigation patterns, social interaction, etc.) (Ge & Ifenthaler, 2017; Ifenthaler, 2015; Loh et al., 2015).

Similarly, Shute (2011) recommends using automated data collection and analysis technologies, i.e., stealth assessment, to reduce educators' efforts by managing learners' artifacts while also gathering evidence of learners' competency states. Further, reflecting on the work of serious games and assessment over a period of ten years, Kim and Ifenthaler (2019) suggested that the field of serious games and assessment can greatly benefit from the application of data analytics, from the design process to classroom implementations. However, it is a challenge to implement data analytics in serious games such that they are grounded in theory and practice, technically sound, and beneficial for learners and educators. Still, recent research has seen advances in serious games embracing data analytics opportunities (Alonso-Fernández et al., 2019).

1.2. Serious games in nursing education

Nursing education plays a pivotal role in preparing future healthcare professionals to meet the evolving challenges in healthcare (Nunn-Ellison et al., 2023). In recent years, significant advancements in nursing education have emerged, such as refined pedagogical approaches, the integration of evidence-based practices into curricula, as well as technological advancements, such as virtual or augmented reality (Mendez et al., 2020) and serious games (Gu et al., 2022). Hence, current advances in nursing education have successfully implemented student-centred and interactive approaches, such as serious games, which foster opportunities for authentic, experiential, and simulated clinical practice scenarios that encourage problem-solving and decision-making skills (Stuckless et al., 2014).

Implementing serious games in nursing education follows various objectives, such as developing communication skills, professional attitudes, ethical and legal awareness, as well as "the development of nursing competency in management of nursing care, clinical reasoning skills, procedural skills, legal practice and quality improvement" (Thangavelu et al., 2022, p. 9). There is a wide variety of serious games use cases in nursing education, including general management of clinical situations (Fonseca et al., 2015), coping with interpersonal conflict situations (Zeffiro et al., 2021), as well as more specific clinical practices, such as cardiopulmonary resuscitation (first aid key survival technique) (Boada et al., 2015) or safe administration of blood transfusion (Tan et al., 2017). Still, Sánchez-Valdeón et al. (2023) document in their recent systematic review that there is little scientific evidence for using serious games in nursing education.

Still, assessment practices in serious games for nursing education strongly focus on external assessments. For instance, Mitchell et al. (2021) investigated the effect of a serious game about influenza among nursing students using a self-report pre- and post-test survey approach. Blanié et al. (2020) explored the impact of a serious game on nursing students' clinical reasoning skills, which were assessed by script concordance tests immediately after playing the game. Another randomised control trial study by Farsi et al. (2021) found knowledge and skills related to cardiopulmonary resuscitation using a knowledge questionnaire and an appraisal checklist. Verkuyl et al. (2017) assessed nursing students' paediatric knowledge, self-efficacy, and satisfaction using standardised (self-report) instruments. Further, Tan et al. (2017) administered multiple (pre- and post-test) instruments to assess participants' knowledge of blood transfusion, the confidence level of the participants in performing the blood transfusion procedure, and performances in the safe administration of the blood product and in responding to a transfusion reaction.

Accordingly, even with the advancements in serious games for nursing education, there is a lack of adaptive assessment methodologies and serious games analytics in practice, as well as a shortage of rigorous evidence concerning the conditions of successful implementation of serious games assessments in nursing education.

1.3. Overview of the present study

The context of the present study is a serious game that functions as an assessment environment for blood transfusion administration among practising nurses of a large health institution. Assessments on core nursing skills, such as blood transfusion administration, are routinely conducted within such health institutions as a regulatory requirement and for quality assurance (Hall et al., 2008). By converting a face-to-face assessment to a serious game, the assessment practice could be standardised, providing learning opportunities for nurses who may not have the chance to experience it in a natural clinical environment and reducing human resources and time required for assessing the nurses at regular intervals. Further, to establish an unobtrusive assessment during game-play (Shute, 2011), the study thought to implement mechanics of embedded assessment enabling real-time adaptation and feedback during individual steps of the serious game (Kovanovic et al., 2023).

Specifically, the study investigates whether gameplay data accurately represents player engagement and activity within a serious game, considering varying gameplay durations and activity levels. Additionally, it examines the ability of an adaptive assessment algorithm to effectively differentiate between player competence levels during gameplay while considering algorithm processing speed. Finally, the research compares the accuracy of the embedded game score and the adaptive assessment algorithm in determining player competence levels. The following research gaps identified guided the three research questions of the present study:

Aster et al. (2024) recently presented a systematic review identifying serious game design elements in medical education. Yet, the findings highlight the absence of assessment elements and a deeper understanding of players' interactions. Hence, our current investigation is unique for this particular blood transfusion serious game and related assessments, as no previous empirical evidence exists. Therefore, our first aim of this study is to investigate whether gameplay data can accurately reflect users' interaction with the serious game and their engagement in related gameplay activities, while considering the

variability in both the duration of gameplay sessions and the scores achieved in game activities.

From a technical perspective, serious games use traditional embedded game scoring mechanisms to identify a player's possible learning progress and outcomes (Kim & Ifenthaler, 2019; Oranje et al., 2019). In contrast, implementations of adaptive assessment algorithms (Spray & Reckase, 1996; Vos & Glas, 2009) and the use of gameplay data (J. Kang, Liu, & Qu, 2017; J. Kang, Liu, & Qu, 2017) highlight the advanced opportunities for fast and accurate decision-making as well as just-in-time feedback in serious games assessments. However, robust findings on the feasibility of embedded game scoring and an exclusively developed adaptive assessment algorithm for identifying learners' expected competence level in this particular blood transfusion serious game are lacking. Hence, our second aim is to determine whether the adaptive assessment algorithm can accurately identify varying levels of user competence throughout gameplay, while taking into account the speed at which the algorithm reaches its final decision.

Last, previous research identified the strengths and weaknesses of traditional game scoring for assessment purposes (Thangavelu et al., 2022) and algorithmic approaches (Loh et al., 2016). Yet, a deeper understanding of different approaches in real-world settings in nursing education is lacking and robust findings for this particular blood transfusion serious game are equally needed. Consequently, our third aim is to provide valuable insights into the relative strengths and weaknesses of each approach and reflect on the relationship between the results obtained from the embedded game score and the adaptive assessment algorithm, particularly in relation to their final decisions regarding the user's competence level.

1.4. Research questions

1. Do gameplay data reflect interaction with the serious game and related gameplay activities, considering the variability in duration of gameplay and game activity scores?
2. Does the adaptive assessment algorithm identify different levels of competence during gameplay, considering the speed of the algorithm's final decision?
3. Is there a relationship in results between the embedded game score and the adaptive assessment algorithm concerning the final decision of competence level?

2. Material and methods

2.1. Blood transfusion serious game

The adminstration of blood products is one of the core competences of a registered nurse (RN) practising in Singapore. Reassessment of of thoses competences is conducted every year within the participating health institution as part of regulatory requirement and for quality assurance.

Traditionally, RNs are required to complete an online learning module, including the theoretical understanding of blood components, blood typing, cross-matching, indications and contraindications for transfusion, risks and complications associated with transfusions, transfusion reactions, and safety protocols, as well as the institution's policy on administering blood and blood products. In addition, nurses are assessed on their competencies in performing safe and accurate blood transfusions administrations, through in-person synchronous assessment involving supervised blood transfusions in clinical settings. Hence, the conduct is resouce intensive, inefficient and challenging as blod transfusions are infrequent in some clinical settings.

Accordingly, blood transfusion administration involves several interrelated phases: blood group cross-matching, patient preparation, blood collection, pre-transfusion and post-transfusion nursing care (Bediako et al., 2021). One of the common and detrimental effects of this life-saving treatment is adverse transfusion reactions, in which preventable human error is the top reason (Lancaster et al., 2021). Hence, the safety and management of blood transfusion largely depend on nurses' knowledge and skills to ensure that the blood administration is coordinated, closely monitored, and adheres to the institution's blood transfusion guidelines (Lancaster et al., 2021).

The *Blood Transfusion Serious Game* (BTSG) follows the institutions's competency standards and competency indicators linked to blood transfusion. Accordingly, the BTSG includes seven stages of game-play, each including multiple actions and required sub-actions to successfully demonstrate competence in blood transfusion (see Fig. 1). The game activities (GA) within the BTSG can be freely selected and repeated at will and there is no time limit within which to complete the game.

Each stage has been assigned a game activity score (GAS), which has been defined and validated by a panel of experienced practitioners regarding the expected competence level of a successful blood transfusion administration. The player could earn a total of 110 GAS. *Stage 1* (20 GAS) asks the player to collect the patient's blood for GXM (Group & Cross Match) investigation. Actions include preparing requisites, identifying the correct patient, collecting the specimen, filling the GXM form, labelling the specimen, and dispatching the blood. *Stage 2* (9 GAS) expects a thorough check of the documents required to proceed with a blood transfusion. Actions include checking for a prescription, checking the GXM and ordering blood, as well as checking the consent form. *Stage 3* (8 GAS) consists of the set-up of the blood box. Actions involve identifying correct objects and preparing the blood box in the correct order. *Stage 4* (3 GAS) expects the player to set up the COW (Computer On Wheels) for blood transfusion. Actions require the placement of an infusion set and kidney dish on the COW. *Stage 5* (16 GAS) involves checking the patient's identity and matching the patient's blood group with the blood unit number. Actions include identifying the correct patient, matching the patient's blood group and blood unit number, and checking the blood bag. *Stage 6* (27 GAS) asks the player to conduct the blood transfusion. Actions include patient education, preparing the patient, and connecting blood to the patient. The final *Stage 7* (27 GAS) involves ending the transfusion and monitoring the patient. Actions required include disconnecting the blood transfusion, checking the patient's vital signs, and performing regular visual inspections on the patient.

Each of the seven stages must be completed in succession to allow the participant to re-enact the correct steps in blood transfusion administration. The participant is deemed to practice competent upon the successful completion of all sven steps in succession. All participants who played the BTSG successfully completed the game eventually and were deemed as practice competent. The analysis of final decision was based on each game play and should not be confused with the actual implementation of BTSG whereby participatns were required to repeat the game play untl successful completeion of all seven steps in succession.

2.2. Data set and participants

There are approximately $N_{RN} = 2200$ registered nurses employed in the health institution who were reassessed for their competency of blood transfusion via the serious game. The data were collected under the health institution's guidelines and principles for human subjects. The initial BTSG data set includes a total of $N_{DR} = 24{,}501$ rows of gameplay actions. An exploratory analysis of the initial BTSG data set identified several data-related issues that needed to be resolved before further analysis, including inconsistencies (e.g., unrealistic gameplay duration, i.e., less than 20 s and more than 1 h of gameplay; no action taken) and missing data. After cleaning the data set, a total of $N_{DR} = 19{,}213$ rows of gameplay actions remained. This final data set includes $N_{RN} = 902$ unique players (i.e., registered nurses).

2.3. Adaptive assessment algorithm

The proposed adaptive assessment algorithm is based on the

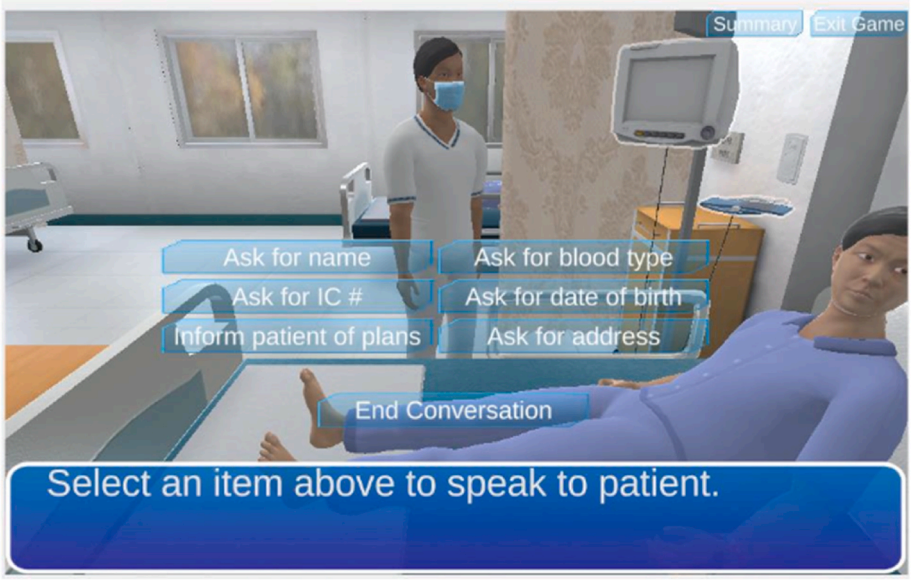


Fig. 1. Screen capture of the Blood Transfusion Game (BTSG).

Sequential Probability Ratio Test (SPRT), a frequently utilised algorithm in Adaptive Mastery Testing (AMT) systems. SPRT follows a decision matrix to decide which one of two hypotheses is more correct (Wald, 1947). Such algorithms enable AMT systems to select and present the most appropriate assessment items to individual learners compared to an expected standard or predefined benchmark (Parshall et al., 2002; Spray & Reckase, 1996). According to Frick (1990), SPRT algorithms are less complex and more practical for implementation as well as require less time for rendering decisions. Implementing the SPRT algorithm requires several steps (SPRT-Step), which shall be outlined in the following (an example is included in the spreadsheet as a separate file). The algorithm's settings can be adjusted to meet more strict or less strict assessment criteria. Such adjustments may be compared in A/B tests to achieve high validity of the competence assessment. Four steps have been applied for implementing the adaptive assessment algorithm: (1) Prior knowledge configuration; (2) Probability settings; (3) Probability ratio; and (4) Final decision. The algorithmic steps are described in detail below.

2.3.1. SPRT-step 1: prior knowledge configuration

Setting the initial configuration of the prior knowledge. If no information about the prior knowledge of a specific competence (e.g., prior knowledge test, prior game-play results) of the learner is available, the probability ratio for correct/incorrect is kept equal (master = .5/non-master = .5) - > Case A ($PriK_{Master\text{-}correct}$ = .5, $PriK_{Master\text{-}incorrect}$ = .5 and $PriK_{Non\text{-}Master\text{-}correct}$ = .5, $PriK_{Non\text{-}Master\text{-}incorrect}$ = .5). In other cases, where the prior knowledge is known, or a previous assessment result of a specific competence is available (e.g., a scenario of the game has been completed), the probability ratio can be adjusted (based on the prior knowledge, e.g., Master = .85/Non-Master = .15) - > Case B ($PriK_{Master\text{-}correct}$ = .85, $PriK_{Master\text{-}incorrect}$ = .15 and $PriK_{Non\text{-}Master\text{-}correct}$ = .40, $PriK_{Non\text{-}Master\text{-}incorrect}$ = .60).

2.3.2. SPRT-step 2: probability settings

This step is iterative for every change in the assessment process, i.e., information from previous assessment states are reflected in the settings. In Example 1 (see spreadsheet), we assume that the assessment result of the first assessment task is incorrect. The *Prior Probability* (PreP) includes the constant value from the prior knowledge configuration (only for the initial question). In other cases, the *Prior Probability* (PreP) consists of the information from the previous assessment state, i.e., *Posterior Probability* (PosP) (see Examples 2, 3, etc.). The *Conditional Probability* (ConP) is obtained from the prior knowledge configuration. The *Joint Probability* (JoiP) is calculated as JoiP = PreP * ConP. The sum of joint probability ($JoiP_{Total}$) is required to calculate the *Posterior Probability* (PosP), i.e., the Sum of JoiP = $\Sigma(JoiP_{Master} + JoiP_{Non\text{-}Master})$. The *Posterior Probability* (PosP) is calculated as $PosP_{Master} = JoiP_{Master}/JoiP_{Total}$ or $PosP_{Non\text{-}Master} = JoiP_{Non\text{-}Master}/JoiP_{Total}$. The *Posterior Probability* (PosP) functions as *Prior Probability* (PreP) in the following sequence of the assessment process (see blue boxes and arrows marking the transition below).

2.3.3. SPRT-step 3: probability ratio

This step determines the *Probability Value* (ProV) and compares it with specified thresholds (i.e., Upper Threshold and Lower Threshold). Based on the comparison, a decision toward the competence achieved is made (see also Step 4): 1. Participant is Master, i.e., no more assessment items needed – competence demonstrated; 2. No decision yet, i.e., additional assessment items required; 3. A participant is Non-Master, i.e., the competence cannot be demonstrated because of too many errors (incorrect answers) made already.

The *Probability Value* (ProV) takes into account the response to the assessment tasks (test item). This can be correct, i.e., the task is solved or incorrect, i.e. the task is not solved. The number of correct (NCR) or incorrect (NIR) responses is accumulated through the iterative steps of the overall assessment sequences. *ProV* is calculated as follows:

$$ProV = ((0,5^{*}(PriK_{Master\text{-}correct})\hat{NCR})^{*}(1\text{-}\ PriK_{Master\text{-}correct})\hat{NIR}))$$
$$/\ ((0,5^{*}(PriK_{Non\text{-}Master\text{-}correct})\hat{NCR})^{*}(1\text{-}\ PriK_{Non\text{-}Master\text{-}correct})\hat{NIR}))$$

The *upper threshold* (UT) and *lower threshold* (LT) are based on the Alpha- and Beta-Error values (following normal distribution). These values can be adjusted in order to meet more strict or less strict assessment criteria. The standard values include Alpha = .025 and Beta = .025. The *upper threshold* is calculated as UT = (1-Beta)/Alpha, e.g., UT = 0,975/0,025 = 39. The *lower threshold* (LT) is calculated as LT = Beta/(1-Alpha), e.g., LT = 0,025/0,975 = 0,02564.

Master is achieved if the *Probability Value* (ProV) is higher than the *Upper Threshold* (UT). If the *Probability Value* (ProV) is between the *Upper Threshold* (UT) and *Lower Threshold* (LT), new assessment items need to be presented. If the *Probability Value* (ProV) is below the *Lower Threshold* (LT), *Non-Master* is determined, and the test can be ended.

2.3.4. SPRT-step 4: final decision (master/non-master)

The final step includes deciding on the competence level after all

assessment items have been solved. Decisions include being Master, i.e., being competent, or Non-Master, i.e., competence not demonstrated per game play. This decision is based on the *Upper Threshold* (UT) obtained after completing the required number of assessment tasks. If the *Probability Value* (ProV) is higher than the *Upper Threshold* (UT), the decision is Master (i.e., competence demonstrated). If the *Probability Value* (ProV) is between *Upper Threshold* (UT) and *Lower Threshold* (LT), a decision cannot be made yet, i.e., an additional assessment task is needed. If the *Probability Value* (ProV) is smaller than *Lower Threshold* (LT), the decision is Non-Master (i.e., competence not demonstrated).

2.3.5. Summary of the applied algorithm

The SPRT algorithm follows a four-step approach, considering the learner's prior knowledge and their performance on assessment tasks. If no prior knowledge is available, the system starts with a neutral assumption about the learner's competence. As the learner completes tasks, the system updates its assessment based on their answers (correct or incorrect). This allows the system to determine if the learner has mastered the competence, needs more assessment, or has not demonstrated the competence. Depending on the application, pre-defined thresholds can be adjusted to trigger the algorithmic decisions.

2.4. Procedure

The health institution, together with a software company, co-developed the BTSG as a competency assessment tool for blood transfusion administration among registered nurses. The game design included subject matter experts from the health industry, software developers, data scientists, and researchers with data science and learning science backgrounds. The game activities were designed to match the above-mentioned competency indicators. Several standardised game scoring metrics were implemented, for example, duration of gameplay, interaction with game objects, or results of game activities. In addition, the adaptive assessment algorithm was implemented to predict whether competence is being demonstrated (Master) or not (Non-Master). For the purpose of this algoirthm, 'competence' refers to game-related data and scores across all seven stages of the BTSG.

Following the piloting of the game prototype, the BTSG was introduced in the health institution as an alternative assessment approach regarding the blood transfusion competence of RNs. The BTSG (see Fig. 1) was implemented on an adopted learning management system on a secure server of the health institution, following industry standards in health education software development (Palominos et al., 2021). Inpatient registered nurses were provided with detailed information about the BTSG and its related assessment approach, as well as a secure link and login credentials to access the BTSG.

After successfully logging into the BTSG, the participants received an introduction related to blood transfusion, the seven stages of game-play, and the gameplay activities required. In addition, participants were informed about the assessment approach and its function as competence alternative competence testing. During gameplay, the BTSG enabled participants to freely select and repeat the game activities (i.e., assessments) at will. The participants received an in-depth debriefing related to their demonstrated competence, including the gameplay data and areas of improvement related to blood transfusion (Roungas et al., 2021).

The data collection covered a period of three months. After concluding the data collection, the data was extracted from the BTSG database and pre-processed by filtering redundant data as well as cleaning the dataset. Following the data-protection practice of the health institution's ethics review committee, all data were stored and analysed using an anonymised procedure. This data-protection practice did not allow for the combination of gameplay data with other person-identifying data collected outside of the BTSG.

2.5. Data analysis

An initial feature selection process was conducted in order to reduce the number of variables in the dataset, i.e., only variables related to the research questions were processed and analysed. From a data science perspective, this feature selection process increases the computational efficiency of statistical procedures and algorithms as well as facilitates the interpretability of the analytics results (Romero et al., 2011). Next, a feature engineering approach was applied (Slater et al., 2017), which helped create new variables required to answer the research questions. The final set of variables are presented in Table 1.

Normal distributions and homogeneity of variances of variables were examined for inferential statistics procedures. All effects were tested at the .05 significance level, and effect size measures were computed where relevant. Data were analysed using r Statistics version 4.3 (https://www.r-project.org) and Python version 3.12 (https://python.org).

2.6. Ethical statement and data availability

This study was approved by the health institutions ethics review committee (*No. 2023/2330*), and participants provided informed consent. The privacy rights of all participants were strictly observed. They were protected by hiding their personal information during the research process. They knew that participation was voluntary and that they could withdraw from the study at any time. There is no potential conflict of interest in this study. The data can be obtained by sending request e-mails to the corresponding author.

3. Results

3.1. Insights from gameplay data (research question 1)

We began our analyses by examining the gameplay data from the BTSG. A total of 19,213 gameplay activities were executed throughout the project period of three months. Participants spent 67,465 min interacting with the BTSG, with an average duration per gameplay activity of $M = 3.51$ min ($SD = 2.95$).

Table 2 shows the gameplay data for each of the seven stages of the BTSG, including the duration per gameplay activity [DUR] and the game activity score [GAS]. The Kolmogorov-Smirnov one-sample tests found significant interindividual differences among the participants in each of the seven stages for the duration of gameplay [DUR] and the game activity score [GAS] (see Table 2).

One-way ANOVA revealed significant differences in duration of gameplay [DUR], $F(6, 19{,}206) = 1459.21$, $p < .001$, $Eta2 = .323$, and for the game activity score [GAS], $F(6, 19{,}206) = 7403.50$, $p < .001$, $Eta2 = .704$, between the different stages (see Table 2 for descriptive statistics). Tukey-HSD test was conducted to determine the source of the differences. As shown in Table 3, only a few pairwise comparisons did not reveal a significant difference between the stages for DUR and GAS.

In addition, the gameplay data included the game result [GAR],

Table 1
Variables in alphabetical order included in the data analyses.

Variable [Abbreviation]	Computation rule	Scale level
Duration [DUR]	timestamp_end – timestamp_start	ratio
Stage [STA]		nominal
game activity score [GAS]	sum of gameplay activity score	ratio
game result [GAR]	required number of competeny indicators achieved	nominal
Adaptive assessment decision [AAD]	number of gameplay activities to reach a decision (Master/Non-Master)	ratio
Adaptive assessment result (AAR)	final decision of SPRT	nominal
Comparison [COM]	similarity between GAR and AAD	nominal

Table 2
Gameplay data separated by stages of the BTSG.

Gameplay data	*BTSG stage (SUM of game activities)*	*M*	*SD*	*Min*	*Max*	*KS-Z*
Duration [DUR]	1 (4546)	258.06	163.19	20	1914	.147***
	2 (2868)	142.34	126.83	20	2664	.186***
	3 (2327)	46.38	43.40	20	1105	.272***
	4 (1642)	61.57	94.69	20	1853	.330***
	5 (2456)	217.32	129.92	20	2238	.144***
	6 (3600)	319.41	187.21	20	2999	.150***
	7 (1774)	323.49	182.36	20	3192	.130***
Game activity score [GAS]	1	14.31	3.24	0	19	.249***
	2	7.48	2.19	0	9	.285***
	3	7.14	1.41	0	8	.393***
	4	2.91	.40	0	3	.526***
	5	12.47	3.28	0	16	.274***
	6	20.66	4.90	0	26	.230***
	7	19.81	7.36	0	27	.308***

Note. DUR = gameplay duration seconds; GAS = points achieved; *KS-Z* = Kolmogorov-Smirnov one-sample test; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 3
Post-hoc analysis of gameplay data separated by stages of the BTSG.

Gameplay data	*Pairwise BTSG stage comparison*					
	1–2	2–3	3–4	4–5	5–6	6–7
Duration [DUR]	$p <$ .001	$p <$ .001	$p =$ .022	$p <$ .001	$p <$ .001	$p =$ .963
Game activity score [GAS]	$p <$ .001	$p =$ .026	$p <$ .001	$p <$ .001	$p <$ .001	$p <$ .001

which denotes if a game activity achieved the required competency indicator. In 46% of game activities, participants achieved the required competence indicators, while 54% of game activities did not demonstrate the required competency indicator. Specifically, 96% of game activities in stage 4 (96%) achieved the required competency indicator, followed by stage 3 (66%), stage 5 (50%), stage 2 (48%), stage 7 (34%), stage 1 (31%), and stage 6 (30%).

A logistic regression was performed to ascertain the effects of duration [DUR] and game activity score [GAS] on the likelihood that the game result will be successful (competency indicator demonstrated) or failure (competency indicator not demonstrated). The logistic regression model was statistically significant, Chi-Square = 2026.38, $df = 2$, $p < .001$. The model explained 13.4% (Nagelkerke R^2) of the variance in game score [GAS] and correctly classified 67.2% of cases. Duration [DUR] and game activity score [GAS] are significant at $p < .001$ (DUR, Wald = 1350.09, $p < .001$; GAS Wald = 490.01, $p < .001$). The odds ratio for DUR is $OR = .99$ (95% CI .99–.99), and for GAS, the corresponding figures are $OR = 1.07$ (95% CI 1.07–1.08).

To sum up, in relation to research question 1, the insights from gameplay data indicate a high number of gameplay activities throughout the different stages of the BTSG. Further, gameplay activities, duration of gameplay, and gameplay score varied significantly between the seven stages of the BTSG. In addition, the game activity score identified almost half of the game activities demonstrating the required competency indicator, while duration and game activity score are significant predictors.

3.2. Performance of the adaptive assessment algorithm (research question 2)

In order to understand the effectiveness of the adaptive assessment algorithm, we investigated its overall performance. It required an average of $M = 6.27$ ($SD = 4.47$) gameplay activities to determine a final algorithmic decision: *Master* (i.e., competence demonstrated) or *Non-Master* (i.e., competence not demonstrated). Additionally, the algorithm could not reach a final decision for some gameplay activities and classified it as *Undecided*.

Table 4 shows the performance of the adaptive assessment algorithm for each of the seven stages of the BTSG. The Kolmogorov-Smirnov one-sample tests found significant interindividual differences among the participants in each of the seven stages of the adaptive assessment decision [AAD] (see Table 4). For stage 1, the AAD indicated that 84% of gameplay activities reached a final decision Master (i.e., competency indicator demonstrated). In comparison, 11% of gameplay activities reached a final decision Non-Master (i.e., competency indicator not demonstrated). For 5% of gameplay activities, no decision (undecided) could be achieved. The AAD decisions for stage 2 to stage 7 have been computed as follows: Stage 2, Master = 70%, Non-Master = 14%, Undecided = 16%; Stage 3, Master = 73%, Non-Master = 3%, Undecided = 24%; Stage 4, Master = 0%, Non-Master = 1%, Undecided = 99%; Stage 5, Master = 74%, Non-Master = 15%, Undecided = 11%; Stage 6, Master = 70%, Non-Master = 30%, Undecided = 0%; Stage 7, Master = 83%, Non-Master = 17%, Undecided = 0%.

One-way ANOVA revealed significant differences in adaptive assessment decision [AAD], $F(6, 19{,}206) = 3039.26$, $p < .001$, $Eta2 = .487$, between the different stages (see Table 4 for descriptive statistics). Tukey-HSD test indicated significant differences in all pairwise comparisons between the stages for AAD, $p < .001$.

A logistic regression was performed to determine the effects of duration [DUR], game activity score [GAS], and adaptive assessment decision [AAD] on the likelihood that the adaptive assessment result [AAR] (final decision of SPRT being Master vs. Non-Master). The logistic regression model was statistically significant, Chi-Square = 1559.54, $df = 3$, $p < .001$. The model explained 15.4% (Nagelkerke R^2) of the variance in adaptive assessment result [AAR] and correctly classified 84.0% of cases. Duration [DUR], game activity score [GAS], and adaptive assessment decision [AAD] are significant at the $p < .001$ (DUR, Wald = 73.39, $p < .001$; GAS Wald = 1220.08, $p < .001$; AAD Wald = 130.51, $p < .001$). The odds ratio for DUR is $OR = .99$ (95% CI .99–.99). For GAS, the corresponding figures are $OR = 1.18$ (95% CI 1.17–1.19), and for AAD, the reported odds ratio is $OR = .93$ (95% CI .92–.94).

To sum up, in relation to research question 2, the adaptive assessment algorithm demonstrated varying levels of effectiveness across different gameplay stages. While it could quickly identify mastery in early stages, its decision-making process became more complex and time-consuming as the game progressed. The algorithm's performance was significantly influenced by factors such as gameplay duration, activity score, and the specific stage of the game.

3.3. Relationships in results between the embedded game score and the adaptive assessment algorithm (research question 3)

Table 5 shows a comparison of decisions concerning the competency indicators for the seven stages between the game result [GAR] and adaptive assessment result [AAR]. The Phi coefficient revealed a significant moderate positive relationship between game result [GAR] and adaptive assessment result [AAR] ($\Phi = .31$, $df =$ [degrees of freedom], p

Table 4
Performance of the adaptive assessment algorithm.

	BTSG stage	M	SD	Min	Max	KS-Z
Adaptive assessment decision [AAD]	1	5.46	2.58	0	20	.441***
	2	4.73	2.37	0	9	.364***
	3	3.84	2.16	0	8	.465***
	4	.03	.29	0	3	.529***
	5	8.59	4.03	0	15	.322***
	6	11.14	4.73	0	26	.197***
	7	6.76	3.24	0	25	.315***

Note. AAD = number of gameplay activities required for final decision; *KS-Z* = Kolmogorov-Smirnov one-sample test; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 5
Cross table showing a comparison of game result [GAR] and adaptive assessment result [AAR] (N = 16,084).

		Adaptive assessment result [AAR]	
		0	1
Game result [GAR]	0	2443	307
	1	6463	6871

Note. 0 = competence not achieved; 1 = competence achieved.

< .001). Out of the 16,084 gameplay activities analysed, identical classifications were found for AAR and GAR as follows: $N = 2443$ (competence not achieved) and $N = 6871$ (competence achieved). However, $N = 6770$ classifications were not identical, indicating that $N = 6463$ classifications of competence achieved by GAR were classified by AAR as competence not achieved. In contrast, $N = 307$ classifications of competence achieved by AAR were classified by AAR as competence not achieved.

To sum up, in relation to research question 3, the performance of the adaptive assessment indicator provides a quick decision regarding the demonstration of a specific competency indicator. Yet, the adaptive assessment and game results vary considerably throughout the seven stages of the BTSG. However, the classification bias appears to stem from the gamer result [GAR], as this indicator classified a considerable number of gameplay activities as having the competence achieved. In contrast, the adaptive assessment results [AAR] classified those gameplay activities as having the competence not achieved.

4. Discussion

This study was conducted to explore the implementation of a serious game for blood transfusion administration and an adaptive assessment algorithm for transforming the conventional face-to-face assessment for practising nurses in a large health institution. The game-embedded assessment mechanics allowed for unobtrusive assessments during gameplay (Shute, 2011). The current findings may further nurture the potential of AI for adaptive competence assessment in serious games. Further, the adaptive assessment algorithm elucidates the viability of assessing learners' expected competence levels throughout various interactions within the game setting (Kovanovic et al., 2023) and provides opportunities for (near) real-time adaptation and feedback mechanisms within the serious game (Loh et al., 2015).

4.1. The findings related to the gameplay data

The analysis of gameplay data from the BTSG has shown several insights into the participants' engagement and performance, as well as demonstrated competency indicators across the seven stages of the serious game. The practising nurses displayed considerable engagement with the BTSG, evidenced by the substantial number of gameplay activities conducted over the three-month project duration. The average interaction time per gameplay activity highlighted a consistent level of involvement across various stages of the BTSG. However, it is essential to note the significant variability in duration per gameplay activity indicated differing levels of immersion or focus among the practising nurses (Divjak & Tomić, 2011; Eseryel et al., 2014). Similarly, the achievement of required competency indicators varied substantially across the seven stages, with stage 4 demonstrating the highest success rate (96%) and stages 6 and 7 showcasing the lowest rates (30–34%). This variance suggests varying levels of complexity of the BTSG (Li et al., 2019). In addition, this variation may also be an effect of the participant's prior knowledge or proficiency (Yang et al., 2021). Further, the logistic regression model provided insights into the factors influencing competency indicator achievement. Both duration (DUR) and game activity score (GAS) emerged as significant predictors. A longer duration per activity was associated with a slightly reduced likelihood of achieving competency indicators (OR = .993), whereas a higher game activity score increased the odds (OR = 1.07). This implies that while spending more time might not necessarily lead to better performance (Carini et al., 2006; Flowerday & Shell, 2015; Ifenthaler et al., 2020), a more productive and engaged gameplay, as reflected in the activity score, positively influences the demonstrated competences.

Accordingly, the findings emphasise the dynamic nature of engagement and performance within the BTSG. Understanding the variability in engagement across game stages and its impact on demonstrated competency indicators is crucial for optimising the serious game design (Ge & Ifenthaler, 2017).

4.2. The findings related to the adaptive assessment algorithm

As outlined in the study, the adaptive assessment algorithm's performance highlights its efficiency in determining competency indicators through gameplay activities. Our results are comparable with previous implementations of the SPRT algorithm (Spray & Reckase, 1996; Vos & Glas, 2009). On average, the implemented adaptive assessment algorithm required approximately six gameplay activities to render a final decision, categorizing the practising nurses as demonstrating the competency indicator or not. This fast decision-making capability is noteworthy, although the presence of undecided classifications in a subset of activities signals areas where the algorithm faced challenges in reaching conclusive determinations (Kandula et al., 2011). The significant variability observed across the seven stages of the BTSG indicates fluctuating levels of demonstrated competency indicators. Hence, the significant interindividual differences among practising nurses at each stage of the game underscore the need for adaptive assessments, which are tailored to individual performance trajectories (Thangavelu et al., 2022).

Noteworthy, the comparison between the game result (GAR) and adaptive assessment result (AAR) highlighted discrepancies in classifying competency indicators. While a moderate positive relationship existed between GAR and AAR, a considerable number of classifications differed between the two approaches. Notably, the adaptive assessment algorithm categorised several gameplay activities of the practising nurses as having the competency indicator not demonstrated, contradicting the embedded game result. This discrepancy suggests a potential bias in the embedded game result towards overestimating competence, while the adaptive assessment algorithm tended to be more conservative in its final decisions.

4.3. Implications

Various implications can be taken from this study's findings that could help advance nursing education further by utilising adaptive serious games assessment. First, the study highlights the adaptive assessment algorithm's strengths in quick competency assessment but also highlights the need for refinement to align more closely with observed gameplay performances. Addressing discrepancies and understanding the sources of bias between assessment methods could enhance the adaptive assessment algorithm's accuracy and reliability in determining competency indicators (Baker & Hawn, 2021). Second, implementing adaptive assessment algorithms requires significant resources, including trained personnel, advanced technology, and possibly time-consuming data analysis. This raises questions about cost-effectiveness and whether these resources could be more beneficially allocated to other initiatives in nursing education (Blanié et al., 2020). Still, advancements in nursing education are continuously shaping the preparation of future nurses, equipping them with the competences needed to navigate the complexities of healthcare (Farsi et al., 2021; Sánchez-Valdeón et al., 2023). Third, while adaptive assessment algorithms aim to elicit the practising nurses' unique strengths and weaknesses, they may inadvertently reinforce a narrow view of learning (Mirata et al., 2020). Accordingly, such adaptive serious games assessment could sideline the importance of the

complexities of nursing education, neglecting critical thinking and social-emotional awareness. As a result, practising nurses may feel pressured to perform well within an adaptive serious games assessment, limiting their individuality and diverse talents (Yue & Jia, 2023). Furthermore, the constant adaptation of assessments might hinder the development of resilience and perseverance in practising nurses (Sánchez De Miguel et al., 2023). In a constantly adapted game environment, practising nurses might miss out on the opportunity to struggle with challenging problems, learn from failures, and build resilience, all of which are crucial for real-world success (Palominos et al., 2021). Fourth, implementing adaptive serious games assessment raises concerns about data privacy and equity (Ifenthaler & Schumacher, 2016; Mutimukwe et al., 2022). These environments often collect vast amounts of data to adapt the learning and assessment experiences. However, this sensitive information can pose risks if not handled ethically or securely (Li et al., 2023; Mutimukwe et al., 2022; Viberg et al., 2022). Last, with the advances of generative AI (GenAI), new opportunities for developing, developing, and implementing serious games arise. Further, game interactions and assessments may be advanced through GenAI, such as text-based responses to player interactions and elaborated feedback throughout gameplay related to the advances of specific competences (Mao et al., 2023). Hence, AI provides a comprehensive array of resources to enhance serious games' design, delivery, and effectiveness. Nevertheless, it is of the utmost importance to carefully consider potential biases and ethical implications to guarantee serious games' inclusivity and educational purpose.

4.4. Limitations

This study is not without limitations. First, the study involves a specific group of registered nurses from a conveniently selected health institution. The findings might not be generalisable to a broader population of nurses or across different healthcare settings, potentially limiting the external validity (Campbell & Stanley, 1963). In addition, due to the strict ethical and data protection regulations, the participants' individual data could not be included and processed in the analyses. This includes the participant's demographic information, for instance, years of practice, job description, or personal characteristics, which may function as co-variates in the analyses conducted (Karami et al., 2017). Second, while serious games aim to replicate real-world scenarios, the game-based learning environment might lack the complexity or nuances of actual clinical settings. This could impact the transferability of competency indicators demonstrated in the BTSG to real-life situations. Third, the generalisability of the adaptive assessment algorithm might be limited. Hence, while the SPRT algorithm appears to be efficient, it might have limitations in capturing all dimensions of competency indicators or might perform differently when applied to different contexts. Alternative data analytics methods may include Bayesian Knowledge Tracing (BKT), which models the probability of a learner's mastery of a competence based on their sequence of responses to items (i.e., game activities). BKT has effectively modelled how learning progresses over time (Scruggs et al., 2023). Another data analytics approach could utilise the Multidimensional Item Response Theory (MIRT), which could handle multiple latent traits (i.e., competency indicators) simultaneously (Kang et al., 2023). Fourth, the data collection period of three months might not capture long-term trends or changes in competency indicators development over extended periods. A longitudinal study duration could offer a more comprehensive perspective of practising nurses' learning trajectories and competence development. Fifth, the presented study did not include a follow-up or verification process, for instance, an outside text to assess if the competency indicators demonstrated in the game translate into improved performance in real clinical settings.

5. Conclusion

The integration of AI and serious games provides a distinctive modus operandi for the investigation of human behaviour and social interactions within a structured and motivating setting. Specifically, adaptive assessment algorithms in serious games offer opportunities for nursing education through tailored tests to individual competences. Accordingly, such approaches present an engaging and scalable environment for evaluating competences across various domains in the healthcare industry. The opportunity to immerse in real-world scenarios, provide immediate feedback, and cater to individual differences makes adaptive serious games assessment a promising avenue for enhancing nursing education. Yet, ongoing research is needed to refine these serious game environments and related adaptive assessment algorithms to ensure their effectiveness, reliability, and validity at acceptable cost as well as under ethically justifiable use of data.

CRediT authorship contribution statement

Dirk Ifenthaler: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Muhittin Sahin:** Writing – review & editing, Data curation. **Ivan Boo:** Writing – review & editing, Project administration. **Darshini Devi Rajasegeran:** Writing – review & editing, Investigation. **Ang Shin Yuh:** Writing – review & editing, Investigation.

Statements on open data, ethics, and conflict of interest

Appropriate permission and ethical approval were obtained for the study (ID: 2023/2330). Informed consent was obtained from all participants, and their privacy rights were strictly observed. To ensure confidentiality, the participants' personal identifiers were removed prior to data processing. The data can be obtained by sending request e-mails to the corresponding author. The authors declare no conflicts of interest.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Andrea Choh, Fazila Aloweni and Liu Kai for their support in conducting the research study. We would also like to thank Serious Games Asia and Singapore General Hospital for their support.

References

Abt, C. C. (1970). *Serious games*. Viking Press.
Abt, C. C. (1987). *Serious games*. University Press of America.
Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education, 141*. https://doi.org/10.1016/j.compedu.2019.103612
Aster, A., Laupichler, M. C., Zimmer, S., & Raupach, T. (2024). Game design elements of serious games in the education of medical and healthcare professions: A mixed-methods systematic review of underlying theories and teaching effectiveness. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-024-10327-1
Aydin, M., Karal, H., & Nabiyev, V. (2023). Examination of adaptation components in serious games: A systematic review study. *Education and Information Technologies, 28*, 6541–6562. https://doi.org/10.1007/s10639-022-11462-1
Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9
Bediako, A. A., Ofosu-Poku, R., & Druye, A. A. (2021). Safe blood transfusion practices among nurses in a major referral center in Ghana. *Advances in Hematology, 2021*, Article 6739329. https://doi.org/10.1155/2021/6739329
Blanié, A., Amorim, M.-A., & Benhamou, D. (2020). Comparative value of a simulation by gaming and a traditional teaching method to improve clinical reasoning skills

necessary to detect patient deterioration: A randomized study in nursing students. *BMC Medical Education, 20*, 53. https://doi.org/10.1186/s12909-020-1939-6

Boada, I., Rodríguez-Benítez, A., García-Gonzalez, J. M., Olivet, J., Carreras, V., & Sbert, M. (2015). Using a serious game to complement CPR instruction in a nurse faculty. *Computer Methods and Programs in Biomedicine, 122*(2), 282–291. https://doi.org/10.1016/j.cmpb.2015.08.006

Boyle, E. A., Hainey, T., Connolly, T. M., Grant, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178–192. https://doi.org/10.1016/j.compedu.2015.11.003

Caballero-Hernández, J. A., Palomo-Duarte, M., & Dodero, J. M. (2017). Skill assessment in learning experiences based on serious games: A systematic mapping study. *Computers & Education, 113*, 42–60. https://doi.org/10.1016/j.compedu.2017.05.008

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education, 47*(1), 1–32. https://doi.org/10.1007/s11162-005-8150-9

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Divjak, B., & Tomić, D. (2011). The impact of game-based learning on the achievement of learning goals and motivation for learning mathematics - literature review. *Journal of Information and Organizational Sciences, 35*(1), 15–30.

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. B. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society, 17*(1), 42–53. https://www.j-ets.net/collection/published-issues/17_1.

Farsi, Z., Yazdani, M., Butler, S., Nezamzadeh, M., & Mirlashari, J. (2021). Comparative effectiveness of simulation versus serious game for training nursing students in cardiopulmonary resuscitation: A randomized control trial. *International Journal of Computer Games Technology, 2021*, Article 6695077. https://doi.org/10.1155/2021/6695077

Flowerday, T., & Shell, D. F. (2015). Disentangling the effects of interest and choice on learning, engagement, and attitude. *Learning and Individual Differences, 40*, 134–140. https://doi.org/10.1016/j.lindif.2015.05.003

Fonseca, L. M., Aredes, N. A., Dias, D., Scochi, C., Martins, J. C. A., & Rodrigues, M. (2015). Serious game e-Baby: Nursing students' perception on learning about preterm newborn clinical assessment. *Revista Brasileira de Enfermagem, 68*, 9–14. https://doi.org/10.1590/0034-7167.2015680102p

Frick, T. W. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research, 6*(4), 479–513.

Ge, X., & Ifenthaler, D. (2017). Designing engaging educational games and assessing engagement in game-based learning. In R. Zheng, & M. K. Gardner (Eds.), *Handbook of research on serious games for educational applications* (pp. 255–272). IGI Global.

Ge, X., Wilson, S. N., Mania Singer, J. T., Thompson, W. M., Kornelson, K. A., Lajos, J., Roper, B., Elizondo, J., Reeder, S. L., Williams, L., & Kleiser, M. L. (2021). The iteration of design and assessment for a digital game to support reasoning in a college algebra course. In C. Aprea, & D. Ifenthaler (Eds.), *Game-based learning across the disciplines* (pp. 273–295). Springer. https://doi.org/10.1007/978-3-030-75142-5_12.

González-Calero, P. A., & Gómez-Martín, M. A. (Eds.). (2011). *Artificial intelligence for computer games*. Springer. https://doi.org/10.1007/978-1-4419-8188-2.

Gu, R., Wang, J., Zhang, Y., Li, Q., Wang, S., Sun, T., & Wei, L. (2022). Effectiveness of a game-based mobile application in educating nursing students on flushing and locking venous catheters with pre-filled saline syringes: A randomized controlled trial. *Nurse Education in Practice, 58*, Article 103260. https://doi.org/10.1016/j.nepr.2021.103260

Hall, L. W., Moore, S. M., & Barnsteiner, J. H. (2008). Quality and nursing: Moving from a concept to a core competency. *Urologic Nursing, 28*(6), 417–425. https://www.proquest.com/scholarly-journals/quality-nursing-moving-concept-core-competency/docview/220151041/se-2.

Ifenthaler, D. (2015). Learning analytics. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 2, pp. 447–451). Sage.

Ifenthaler, D., Eseryel, D., & Ge, X. (2012). Assessment for game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning. Foundations, innovations, and perspectives* (pp. 3–10). Springer. https://doi.org/10.1007/978-1-4614-3546-4_1.

Ifenthaler, D., Gibson, D. C., & Zheng, L. (2020). Attributes of engagement in challenge-based digital learning environments. In P. Isaias, D. G. Sampson, & D. Ifenthaler (Eds.), *Online teaching and learning in higher education* (pp. 81–91). Springer. https://doi.org/10.1007/978-3-030-48190-2_5.

Ifenthaler, D., & Schumacher, C. (2016). Student perceptions of privacy principles for learning analytics. *Educational Technology Research & Development, 64*(5), 923–938. https://doi.org/10.1007/s11423-016-9477-y

Jerčić, P., & Sundstedt, V. (2019). Practicing emotion-regulation through biofeedback on the decision-making performance in the context of serious games: A systematic review. *Entertainment Computing, 29*, 75–86. https://doi.org/10.1016/j.entcom.2019.01.001

Kandula, S., Ancker, J. S., Kaufman, D. R., Currie, L. M., & Zeng-Treitler, Q. (2011). A new adaptive testing algorithm for shortening health literacy assessments. *BMC Medical Informatics and Decision Making, 11*, 52. https://doi.org/10.1186/1472-6947-11-52

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior, 72*, 757–770. https://doi.org/10.1016/j.chb.2016.09.062

Kang, C., Liu, N., Zhu, Y., Li, F., & Zeng, P. (2023). Developing College students' computational thinking multidimensional test based on Life Story situations. *Education and Information Technologies, 28*, 2661–2679. https://doi.org/10.1007/s10639-022-11189-z

Karami, A., Farokhzadian, J., & Foroughameri, G. (2017). Nurses' professional competency and organizational commitment: Is it important for human resource management? *PLoS One, 12*(11), Article e0187863. https://doi.org/10.1371/journal.pone.0187863

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In D. Ifenthaler, & Y. J. Kim (Eds.), *Game-based assessment revisted* (pp. 3–12). Springer. https://doi.org/10.1007/978-3-030-15569-8_1.

Kitchen, G. B., & Humphreys, J. (2014). Serious games in medical education. Do they have a role in anaesthetic training. *Trends in Anaesthesia and Critical Care, 4*(2–3), 63–66. https://doi.org/10.1016/j.tacc.2014.03.001

Kovanovic, V., Azevedo, R., Gibson, D. C., & Ifenthaler, D. (2023). Data for unobtrusive observations of learning: From trace data to multimodal data. In V. Kovanovic, R. Azevedo, D. C. Gibson, & D. Ifenthaler (Eds.), *Unobtrusive observations of learning in digital environments. Examining behavior, cognition, emotion, metacognition and social processes using learning analytics* (pp. 119–121). Springer. https://doi.org/10.1007/978-3-031-30992-2_8.

Laamarti, F., Eid, M., & El Saddik, A. (2014). An overview of serious games. *International Journal of Computer Games Technology, 358152*. https://doi.org/10.1155/2014/358152

Lancaster, E., Rhodus, E., Duke, M., & Harris, A. (2021). Blood transfusion errors within a health system: A review of root cause analyses. *Patient Safety, 3*(2), 78–91. https://doi.org/10.33940/med/2021.6.6

Li, W., Funk, M., Li, Q., & Brombacher, A. (2019). Visualizing event sequence game data to understand player's skill growth through behavior complexity. *Journal of Visualization, 22*, 833–850. https://doi.org/10.1007/s12650-019-00566-5

Li, F., Ruijs, R., & Lu, Y. (2023). Ethics & AI: A systematic review on ethical concerns and related strategies for designing with ai in healthcare. *AI, 4*(1), 28–53. https://doi.org/10.3390/ai4010003

Loh, C. S., Li, I.-H., & Sheng, Y. (2016). Comparison of similarity measures to differentiate players' actions and decision-making profiles in serious games analytics. *Computers in Human Behavior, 64*, 562–574. https://doi.org/10.1016/j.chb.2016.07.024

Loh, C. S., & Sheng, Y. (2014). Maximum similarity index (msi): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior, 39*, 322–330. https://doi.org/10.1016/j.chb.2014.07.022

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). Serious games analytics: Theoretical framework. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics. methodologies for performance measurement, assessment, and improvement* (pp. 3–29). Springer. https://doi.org/10.1007/978-3-319-05834-4_1.

Mao, J., Chen, B., & Liu, J. C. (2023). Generative artificial intelligence in education and its implications for assessment. *TechTrends, 68*, 58–66. https://doi.org/10.1007/s11528-023-00911-4

Mendez, K. J. W., Piasecki, R. J., Hudson, K., Renda, S., Mollenkopf, N., Smith Nettles, B., & Han, H.-R. (2020). Virtual and augmented reality: Implications for the future of nursing education. *Nurse Education Today, 93*, Article 104531. https://doi.org/10.1016/j.nedt.2020.104531

Mirata, V., Hirt, F., Bergamin, P., & van der Westhuizen, C. (2020). Challenges and contexts in establishing adaptive learning in higher education: Findings from a delphi study. *International Journal of Educational Technology in Higher Education, 17*, 32. https://doi.org/10.1186/s41239-020-00209-y

Mitchell, G., Leonard, L., Carter, G., Santin, O., & Brown Wilson, C. (2021). Evaluation of a 'serious game' on nursing student knowledge and uptake of influenza vaccination. *PLoS One, 16*(1), Article e0245389. https://doi.org/10.1371/journal.pone.0245389

Mutimukwe, C., Viberg, O., Oberg, L.-M., & Cerratto-Pargman, T. (2022). Students' privacy concerns in learning analytics: Model development. *British Journal of Educational Technology, 53*(4), 932–951. https://doi.org/10.1111/bjet.13234

Nunn-Ellison, K., Tillson, M., Ard, N., & Farmer, S. (2023). Assessment and evaluation: Nursing education and ACEN accreditation. *Teaching and Learning in Nursing, 18*(4), 457–462. https://doi.org/10.1016/j.teln.2023.06.009

Oranje, A., Mislevy, B., Bauer, M. I., & Jackson, G. T. (2019). Summative game-based assessment. In D. Ifenthaler, & Y. J. Kim (Eds.), *Game-based assessment revisted* (pp. 37–65). Springer. https://doi.org/10.1007/978-3-030-15569-8_3.

Palominos, E., Levett-Jones, T., Power, T., Alcorn, N., & Martinez-Maldonado, R. (2021). Measuring the impact of productive failure on nursing students' learning in healthcare simulation: A quasi-experimental study. *Nurse Education Today, 101*, Article 104871. https://doi.org/10.1016/j.nedt.2021.104871

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer.

Ravyse, W. S., Seugnet Blignaut, A., Leendertz, V., & Woolner, A. (2017). Success factors for serious games to enhance learning: A systematic review. *Virtual Reality, 21*, 31–58. https://doi.org/10.1007/s10055-016-0298-4

Ritterfeld, U., Cody, M., & Vorderer, P. (2009). Introduction. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games. Mechanisms and effects* (pp. 3–9). Routledge. https://doi.org/10.4324/9780203891650.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. CRC Press.

Roungas, B., Meijer, S., & Verbraeck, A. (2021). The tacit knowledge in games: From validation to debriefing. In M. Wardaszko, S. Meijer, H. Lukosch, H. Kanegae, W. C. Kriz, & M. Grzybowska-Brzezińska (Eds.), *Simulation gaming through times and*

disciplines (Vol. 11988, pp. 74–83). Springer. https://doi.org/10.1007/978-3-030-72132-9_7. *ISAGA 2019*.

Sánchez De Miguel, M., Ortiz de Elguea, J., Gómez-Gastiasoro, A., Urcola, F., Gema Cid-Expósito, M., Torres-Enamorado, D., & Orkaizagirre-Gomara, A. (2023). Patient safety and its relationship with specific self-efficacy, competence, and resilience among nursing students: A quantitative study. *Nurse Education Today, 121*, Article 105701. https://doi.org/10.1016/j.nedt.2022.105701

Sánchez, J., & Olivares, R. (2011). Problem solving and collaboration using mobile serious games. *Computers & Education, 57*(3), 1943–1952. https://doi.org/10.1016/j.compedu.2011.04.012

Sánchez-Valdeón, L., Casado-Verdejo, I., Barrionuevo, L., Fernández-Martínez, E., Liébana-Presa, C., Pereira, R., & Gomes, L. (2023). Implementation of serious games in nursing student education: A systematic review. *Teaching and learning in nursing*. https://doi.org/10.1016/j.teln.2023.08.015

Scruggs, R., Baker, R. S., Pavlik, P. I., McLaren, B. M., & Liu, Z. (2023). How well do contemporary knowledge tracing algorithms predict the knowledge carried out of a digital learning game? *Educational Technology Research & Development, 71*, 901–918. https://doi.org/10.1007/s11423-023-10218-z

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Information Age Publishers.

Slater, S., Joksimović, S., Kovanović, V., Baker, R. S., & Gašević, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics, 42*(1), 85–106. https://doi.org/10.3102/1076998616666808

Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A meta-analysis of data collection in serious games research. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics. methodologies for performance measurement, assessment, and improvement* (pp. 31–55). Springer. https://doi.org/10.1007/978-3-319-05834-4_2.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405–414. https://doi.org/10.2307/1165342

Squire, K. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19–29. https://doi.org/10.3102/0013189X035008019

Stuckless, P., Hogan, M., & Kapralos, B. (2014). Virtual simulations and serious games in community health nursing education: A review of the literature. In M. Ma, L. Jain, & P. Anderson (Eds.), *Virtual, augmented reality and serious games for health* (pp. 145–158). Springer. https://doi.org/10.1007/978-3-642-54816-1_8.

Tan, A. J. Q., Lee, C. C. S., Lin, P. Y., Cooper, S., Lau, L. S. T., Chua, W. L., & Liaw, S. Y. (2017). Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: A randomized controlled trial. *Nurse Education Today, 55*, 38–44. https://doi.org/10.1016/j.nedt.2017.04.027

Thangavelu, D. P., Tan, A. J. Q., Cant, R., Chua, W. L., & Liaw, S. Y. (2022). Digital serious games in developing nursing clinical competence: A systematic review and meta-analysis. *Nurse Education Today, 113*, Article 105357. https://doi.org/10.1016/j.nedt.2022.105357

Tolks, D., Schmidt, J. J., & Kuhn, S. (2024). The role of AI in serious games and gamification for health: Scoping review. *JMIR Serious Games, 12*, Article e48258. https://doi.org/10.2196/48258

Tsopra, R., Courtine, M., Sedki, K., Eap, D., Cabal, M., Cohen, S., Bouchaud, O., Mechaï, F., & Lamy, J.-B. (2020). AntibioGame®: A serious game for teaching medical students about antibiotic use. *International Journal of Medical Informatics, 136*, Article 104074. https://doi.org/10.1016/j.ijmedinf.2020.104074

Verkuyl, M., Romaniuk, D., Atack, L., & Mastrilli, P. (2017). Virtual gaming simulation for nursing education: An experiment. *Clinical Simulation in Nursing, 13*(5), 238–244. https://doi.org/10.1016/j.ecns.2017.02.004

Viberg, O., Engström, L., Saqr, M., & Hrastinski, S. (2022). Exploring students' expectations of learning analytics: A person-centered approach. *Education and Information Technologies, 27*, 8561–8581. https://doi.org/10.1007/s10639-022-10980-2

Vlachopoulos, D., & Makri, A. (2017). The effect of games and simulations on higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education, 14*, 22. https://doi.org/10.1186/s41239-017-0062-1

Vos, H. J., & Glas, C. A. W. (2009). Testlet-based adaptive mastery testing. In W. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing. Statistics for social and behavioral sciences* (pp. 389–407). Springer. https://doi.org/10.1007/978-0-387-85461-8_20.

Wald, A. (1947). *Sequential analysis*. John Wiley.

Yang, X., Rahimi, S., Shute, V., Kuba, R., Smith, G., & Alonso-Fernández, C. (2021). The relationship among prior knowledge, accessing learning supports, learning outcomes, and game performance in educational games. *Educational Technology Research & Development, 69*, 1055–1075. https://doi.org/10.1007/s11423-021-09974-7

Yue, Y., & Jia, Y. (2023). Fear of negative evaluation: A cross-sectional study among undergraduate nursing students. *Nurse Education Today, 121*, Article 105678. https://doi.org/10.1016/j.nedt.2022.105678

Zeffiro, V., Di Fuccio, R., Vellone, E., Alvaro, R., & D'Agostino, F. (2021). Serious game and negotiation skills in nursing students: A pilot study. In Z. Kubincova, L. Lancia, E. Popescu, M. Nakayama, V. Scarano, & A. Gil (Eds.), *Methodologies and intelligent systems for technology enhanced learning* (pp. 91–98). Springer. https://doi.org/10.1007/978-3-030-52287-2_9.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25–32. https://doi.org/10.1109/MC.2005.297, 38(9), 25–32.