https://doi.org/10.1109/SeGAH57547.2023.10253804

Serious Games Assessment: Analytics, Measurement, and Visualization of Nursing Competencies

Christian S. Loh
School of Education
Southern Illinois University
Carbondale, IL, USA
csloh@siu.edu

Yanyan Sheng
Social Sciences Division
University of Chicago
Chicago, IL, USA
y.sheng@uchicago.edu

Darshini Devi d/o Rajasegeran
Singapore General Hospital
Singapore
darshini.devi.rajasegeran
@sgh.com.sg

Kai, Liu
Serious Games Asia
Singapore
liukai@seriousgamesasia.com

Andrea Chau Lin, Choh
Singapore General Hospital
Singapore
andrea.choh.c.l@sgh.com.sg

Shin Yuh, Ang
SingHealth Nursing Group
SGH, Singapore
ang.shin.yuh@singhealth.com.sg

***Abstract*—It is crucial for practicing nurses to have comprehensive knowledge on the safety and management of blood transfusion administration (BTA), which is a common core competency for registered nurses worldwide. Skill competency assessments are regularly conducted at healthcare institutions to ensure practicing nurses meet standards. However, direct (face-to-face) observations of nursing competencies have many inconsistency issues, including quality of assessment (personal judgment), suitability in matching cases (due to complications), manpower and time availability (among assessor, nurses, and cases). The advent of serious games analytics makes it possible to deploy serious games as remote (psychometric) assessment tools.**

This study assessed the (psychometric) validity and reliability of a Blood Transfusion Serious Game (BTSG) as an assessment tool for measuring nursing competencies in blood transfusion administration. Empirical data (serious games analytics) were collected *in situ* the game. We evaluated the following psychometric properties, namely: action analysis (using difficulty and discrimination measures), construct validity (through an exploratory factor analysis with principal factor analysis and a Promax rotation), and internal-consistency reliability (via Cronbach's alpha). After removing game actions that were not very discriminating or did not load on a single factor, we found the resulting game to be both internally reliable and valid in measuring six remotely related latent subconstructs. Suggestions to visualize the analytics as individual and group competency profiles were provided to assist management and administration with recommendations for the (re)training of underperformed nurses (for remediation or probation) and improvement of the measurement for future serious games assessment.

Keywords—serious games assessment, nursing, competency assessment, measurement, factor analysis, blood transfusion

I. INTRODUCTION

When assessment and serious games are mentioned together, people commonly envisioned some kind of scoring system (usually in the form of a dashboard, with/without feedback) to inform trainees of their performance within the simulated training environment (i.e., serious games). More recently, psychometric and measurement researchers have become increasingly interested in the area of serious games assessment (or, game-based assessment) [1], [2]. This direction of inquiry can be beneficial for healthcare research, where skill competency assessments take place regularly to ensure standards of performance are being met. This study investigates whether serious games analytics collected *in situ* a Blood Transfusion Serious Game (BTSG) can meet psychometric validity and reliability requirements as an assessment tool for nursing competency in blood transfusion administration.

A. Blood Transfusion Administration

Registered nurses who wish to practice in Singapore must meet standards for the core competencies set by the Singapore's Nursing Board [3]. Nurses must be able to demonstrate their knowledge and acquired procedural skills [4] in these regularly conducted competency assessments (testing exercises), to meet professional standards [5] for regulatory requirement and quality assurance. Those failing to meet the competency standards may face consequences, including employment probation.

One such core competency is that of Blood Transfusion Administration (BTA), which consisted of five interrelated phases, namely: (1) group and cross-match of blood, (2) patient preparation, (3) blood collection, (4) pre-transfusion, and (5) post-transfusion nursing care [6]. Because human errors [7] have been identified as the top factor of adverse (blood) transfusion reaction, strict guidelines have been put in place by healthcare institutions to closely monitor and ensure that nurses adhere to BTA procedures [7], [8]. This kind of competency assessment is a common practice for the nursing profession.

Current BTA competency assessment involves completing an online training module, followed by a direct (in-person) observation by an assessor who is collocated with the nurse(s) in the same department [9]. However, as the frequency of blood transfusion can differ greatly from department to department, this often resulted in scheduling limitations for when nurses may be assessed. Moreover, competency assessment performed by human assessor(s) can lack consistency due to variations in personal judgment. The variation from case to case and the complications that can arise in blood transfusion further

978-1-6654-8439-8/22/$31.00 ©2023 IEEE

complicate the matter of using real-life clinical cases for nursing competency assessment.

Hence, the nursing department of the Singapore General Hospital (SGH) decided to develop and pilot-test a BTSG to be a cross departmental competency assessment platform. If proven viable, the serious games assessment may even be deployed to a wider group of medical workers.

B. Motivation

A BTSG for nursing competency assessment can have the following perceivable benefits:

- It can relieve the departments that are strapped with limited opportunities for BTA assessment, especially when in-person meetings become impossible: say, during the pandemic.
- Time and manpower required to schedule for assessment can be greatly reduced.
- Case to case variation can be completely eliminated as the same cases/scenarios will be used for all departments.
- The game can be designed to show *only* regular cases for competency assessment, unlike real-life cases, which may contain complications that nurses have yet to encounter.
- The quality of assessment will be 'standardized' as it no longer depends on the assessors' personal judgment.
- If proven viable, the assessment platform can be expanded to include other types of healthcare procedures and competency assessments; thereby reducing costs of development for affiliated departments and hospitals.
- Should the needs arise, new cases (such as complications that are currently beyond the core competencies) may be added to the serious game as advanced assessment.

Empirical data generated by the nurses *in situ* the game was collected as serious games analytics [10]. This large-scale study examined if the current sets of analytics collected are sufficient to measure the nursing competency in BTA. Psychometric properties of BTSG were assessed using item analysis, while the reliability and construct validity of the game was evaluated through estimating the internal-consistency reliability and establishing evidence on internal structure. Individual nurses' performance analytics were processed and graphically visualized to serve as evidence of passing the assessment (i.e., meeting the standards), or requiring retraining or probation.

II. LITERATURE REVIEW

Serious games are digital games created not for the primary purpose of entertainment, but for education, training, and problem-solving [10]. Good serious games should include assessment components and analytics to empirically measure training outcomes and assess users' performance [11], [12]. Serious games created for healthcare training can enable the players to be engaged in simulated medical environments and situations, and negate the need for providing facilities and resources, or face-to-face assessment [13]. Applications of serious games in healthcare are plenty, ranging from behavior modification (e.g., learning of procedural skills or improving communications among doctors and nurses) to scenario-based training (e.g., patient simulation), and others [14].

Increasingly, serious games researchers are not limited to using analytics to generate feedback for individual performances, but to regard serious games as valid *assessment tools* for measuring competencies.

A. Serious Games as Training and Assessment Tool

The *Serious Games Initiatives* first debuted in year 2002 [15]. A literature review of serious games research over the past 20 years indicated that sensory stimuli, gameplay and challenges embedded in serious games were able to stimulate players' motivation and improve learning engagement [12]. Serious games for healthcare (professional training) can further provide players with the opportunity to (a) learn and strengthen theoretical knowledge through (medical) simulation with intuitive feedback, (b) repeatedly practice without waiting for a case to become available, and (c) be assessed consistently and without an assessor to be physically present [16].

One would think that, after all these years, healthcare management and administration would readily embrace serious games for professional training and assessment. However, while the number of nursing-related research with serious games is clearly on the rise, most of these studies remain steeped in curriculum supplements for use in nursing schools [16]–[18], rather than for professional training and skills improvement [19]. It appears to be easier for schools to justify an experimental research using serious games than for institutions to warrant a high-cost serious games development for large-scale training. The reason is that serious games (still) lack well-designed assessment with empirical evidence that can point to clear training needs and convincing benefits [20]. Without such evidence, it would be difficult for the management and administration to justify the cost of development to shareholders.

The issue is really *not* a lack of serious games research, as there are myriads of serious games studies and game-based research within the healthcare literature. The problem associated with serious games assessment in healthcare research is the amount of *noise* created by inappropriate research designs within the literature. This noise in the literature is interfering with the clear signals (convincing benefits) needed by management and administration, and until researchers understand what not to do, the problem will continue to exist.

B. Noise in Current Research Caused by Media Comparisons

Let us first explain what this 'noise' is within the current serious games literature. Many medical researchers approach serious games like a medical intervention (e.g., drugs used in clinical trials). They approached effective assessment of serious games in healthcare by way of the "gold standard of medical research", i.e., Randomized Controlled Trials (RCTs) [21]. While RCT is desirable for treatment/intervention comparison research – because researchers knew a great deal about how drugs are metabolized, for instance -- the between-groups comparison method can involve confounding bias when used in *media* comparisons. This is because serious games are a type of media, and it is a fallacy to think serious games (for cognitive

learning) can be equated to medical interventions (e.g., drug metabolism), which is a *false equivalence*.

While RCTs maybe the gold-standard methodology for the comparison study of drug effectiveness in medical intervention, it serves little purpose in clarifying how people learn with serious games. No researcher knows for sure, and the literature remains equivocal, about how any media – not limited to serious games, may be learned effectively. The vast number of variables that can affect how people learn means the interpretation of research outcome can be fraught with confounders (covariates, in statistical terms). Correlation is not causation!

Media comparisons first gained notoriety in education research during the 1980s [22], [23] when researchers attempted to compare the efficacy of one type of media (such as computer-based instruction, game-based learning, or some technology) against another (say, a controlled group, a different technology, even traditional classroom teaching). Due to the false equivalency mentioned above, media comparison studies often make use of between-groups comparisons (also known as pre-test/post-test designs, just like RCTs) to study the effects of game-based learning. Confounders in media comparisons exist because there are factors that are either unknown, or cannot be accounted for (hence, often not acknowledged). Interested readers are referred to the seminal papers on media comparisons by Richard Clark [23]–[25].

In healthcare research with serious games, the knowledge learned, interpretation, and the application of gameplay must be verified to be reliable, valid, and specific [14]. However, the continual (mis)use of media comparisons (and RCTs) in healthcare serious games can sometimes lead to 'inconclusive outcomes' (i.e., *noise* that pollutes the clear signal) that interferes with the empirical evidence needed by the administration and management to invest in the technology. [See this systematic review [26] where some conflicting results (i.e., *noise*) are caused by studies with flawed RCTs designs.]

Furthermore, because RCTs require larger sample sizes and are more expensive to conduct, the presence of confounders in the study can cause a spurious association, and introduce errors in hypothesis specification – often, in the form of *non-significant findings*. We must caution readers to never interpret "non-significant difference" between treatment (with serious games) and the control (in RCTs) to mean that both media are *just as good*. The prevalence of the 'just-as-good' fallacy in the literature underscore not only the presence of the false equivalency in media comparisons, but also why over 800 scientists are now calling for the 'retirement of statistical significance' [27]!

Churchill once said, "Those that fail to learn from history are doomed to repeat it." Despite multiple calls to abandon media comparison research [22], [25], [28], [29], such studies can still be found in recent research and with periodic resurgence among younger researchers who are not aware of its history. Even though not every RCT is flawed, flawed RCTs with *false equivalency* can produce inconclusive results that crept into the literature as *noise*. [A discussion of further *noise* removal in serious game research is beyond the scope of this paper, but is available [29, p. 34] for readers who are interested.]

Hence, our recommendation to the healthcare researchers is to consider alternative research designs (other than RCTs) that are more appropriate for the assessment of healthcare knowledge, skills, and techniques learned, and to evaluate if transfer of learning indeed occur from games to practice. (We provide two suggestions in the textbox below.)

Two Research Alternatives to RCTs (Between-Group Comparisons) for Serious Games Research:

(a) Within-Group Comparisons (also, Repeated Measures). A within-group comparison using repeated measures involves more statistical power than RCTs. Furthermore, there is no need for a separate control group because the participants themselves also serve as their own control.

(b) Serious Games Analytics – by way of *in situ* data collection (using a database server) is the superior option for serious game research. Researchers can collect user-generated data directly *in situ* the games, interpret them into analytics, and visualized it as evidence for performance, measurement, and assessment [18].

C. Serious Games as Assessment Tools

In recent years, healthcare research has increasingly focused on using serious games as competency assessment tools for healthcare-related training [1], [30], [31] and evaluating their psychometric properties. For example, [32]–[34] all examined the (measurement) validity evidence of healthcare serious games for training assessment.

Given the plethora of standardized protocols and training/competency requirements in healthcare and nursing where patient safety takes precedence, it can be very useful if serious games can serve both functions of healthcare training and competency assessment. For latter purposes, and especially when numeric scores are used to represent nurse competencies after a gameplay, two fundamental psychometric aspects that need to be considered, as in traditional educational and psychological measurement, are reliability and validity. While classical test theory (CTT) [35] serves as the foundation for reliability, and especially internal-consistency reliability to ensure that game actions developed for assessment purposes are consistent in measuring the same underlying latent competency, it is important that healthcare researchers develop validity evidence necessary to provide support of game use and inferences about individual players based on scores.

Since the 1999 Standards for Educational and Psychological Testing [36], construct validity has been positioned as the overarching consideration of validity with various types of evidence supporting the validity argument, including content, response processes, internal structure, relations to other constructs, and consequences of testing. Among them, validity evidence on internal structure, i.e., the extent of how the relationship between game actions and components reflect the construct, has been important and commonly gathered using factor analysis [37].

In spite of the advancement of psychometric theories, these methods have been slow to be applied to healthcare research, especially those using serious games as assessment tools for nurse competency.

III. METHODOLOGY

A serious game for BTA was developed based on actual clinical procedures for core competency assessment by the

nursing department of Singapore General Hospital. The BTSG (Blood Transfusion Serious Game) comprised a total of seven stages, with 3-32 game actions per stage, covering specific training objectives in blood transfusion – totaling 107 game actions (see Table 1 for a summary). The content validity of the blood transfusion procedure presented in the game was established by four external subject-matter experts.

A total of 1093 registered nurses took part in this study. Prior to analysis, a data cleaning step was performed to remove actions with no valid response (1 action was found in Stage 6). The data cleaning process yielded 1093 complete observations (nurses) with responses to 106 game actions. Each player's response was coded as 0 (no attempt or wrong), .5 (correct but not in correct sequence), and 1 (correct).

TABLE 1: GAME OBJECTIVES AND ACTIONS FOR EACH STAGE OF BTSG

Stage	Game Objectives	Total Actions
1: Collect Blood	Identify patient, collect blood specimen	17
2: Order Blood	Check group and cross-match, order blood	9
3: Prep Blood	Prepare blood box in the correct order	8
4: Prep Cart	Ready instrument and infusion set	3
5: Verify Match	Check blood ordered is a correct match to patient	16
6: Begin Transfusion	Patient education, prepare patient and blood, start transfusion	27
7: Monitor Transfusion	Monitor the process, check on patient, end/stop transfusion	27

IV. DATA ANALYSIS

The psychometric properties of the assessment in BTSG were assessed using the sample data via the following aspects:

- Each game action was analyzed using CTT, where action difficulty and discrimination indices were obtained to evaluate the properties of individual actions.
- Construct validity was evaluated using Exploratory Factor Analysis (EFA) to determine the game assessment's internal structure.
- Internal consistency reliability was estimated using Cronbach's alpha and its 95% confidence interval (CI) for each stage of the competence measured.

A. Game Action Analysis

Given the player response was coded 0-1, game action difficulty can be obtained by computing the proportion of correct responses (or action mean), and action discrimination can be obtained by computing the correlation between the action score and total game score while removing the respective action score (corrected action-total correlation).

Similar to item difficulty and discrimination indices, larger values on action difficulties indicate higher proportions of correct responses and, therefore, easier actions, whereas larger values on action discriminations indicate their higher abilities in discriminating between high performing vs. low performing game players. Typically, we look for game actions with medium difficulty levels (with proportion correct ranging from .25 to .85) and with positive discriminations above a certain threshold. In this study, game actions flagged with low/high difficulty levels were not considered to be a problem. As a matter of fact, it is reasonable to assume action difficulty levels to range from 0-1 in serious games assessments like the BTSG. Only game actions flagged with discriminations lower than .10 were not desirable and hence removed from further analyses.

B. Exploratory Factor Analysis

For construct validity, EFA was carried out using principal factor analysis where inter-game–action polychoric correlations were analyzed. The number of factors to extract was initially determined by Kaiser's rule, scree plot, parallel analysis [26], as well as the Minimum Average Partial (MAP) criterion [27]. To achieve a simple structure, a Promax rotation was used to iteratively remove the following:

- Game actions that did not load high (using a cutoff of .35) on any factor,
- Game actions that cross-loaded, or
- Game actions that loaded high on a factor that differs from the rest of the game actions in the same stage.

V. RESULTS

A. Game Action Analysis

For the total of 106 game actions designed to measure the objectives in the seven stages, action difficulties and discriminations were obtained as the means and corrected action-total correlations, and summarized in Table 2. As noted earlier, we look for game actions with medium difficulty levels (with proportion correct ranging from .25 to .85) and with positive discriminations that are larger than .10. Values outside this range are flagged in parentheses.

For example, game action 5_4 (action number 4 in stage 5) had a difficulty of .275 that is within the range but a negative discrimination (-.049), indicating that high-performing game players were less likely to respond to this action correctly compared with low-performing game players, and therefore is not desirable. On the other hand, game action 1_1 (action number 1 in stage 1), with a discrimination index of .359, discriminated well although its difficulty level was relatively high, indicating an easier action. To allow for a wider range of action difficulties, we examined and removed ten actions with discrimination indices lower than .10 (highlighted in both bold and italics in Table 2) from further analyses.

B. Construct Validity and Internal Consistency Reliability Evidence

With the remaining game actions, a total of 8 factors were initially extracted using principal factor analysis, as suggested by the MAP criterion. A simple structure with a six-factor solution was obtained after removing 14 game actions that either did not load high on any of the factors, loaded high on more than one factor, or loaded high on other factors.

The resulting internal structure of the BTSG is summarized in top two panels of Table 3, where it is clear that the six extracted factors contained originally designed game actions for stages 1, 2, 3, 5, 6, and 7, respectively.

TABLE 2. DIFFICULTY AND DISCRIMINATION FOR EACH GAME ACTION (N = 1093); ACTIONS WHERE DISCRIMINATIONS < .1 (IN BOLD AND ITALICS) WERE REMOVED FROM FURTHER ANALYSIS.

Game Action	Difficulty	Discrimination	Game Action	Difficulty	Discrimination
1_1	(.855)	.359	6_1	.782	.328
1_2	.835	.357	6_2	(.876)	.325
1_3	(.937)	.214	6_3	.787	.262
1_4	.697	.375	***6_4***	***(.210)***	***(-.010)***
1_5	***(.185)***	***(.001)***	***6_5***	***(.172)***	***(-.030)***
1_6	.824	.317	6_6	(.888)	.282
1_7	***.363***	***(.099)***	6_7	.668	.231
1_8	.811	.298	6_8	(.935)	.429
1_9	.581	.381	6_9	(.931)	.435
1_10	.521	.426	6_10	(.904)	.452
1_11	.594	.414	6_11	.664	.514
1_12	***.319***	***(.013)***	6_12	.727	.426
1_13	***(.070)***	***(.037)***	6_13	.755	.521
1_14	.785	.376	6_14	.840	.471
1_15	(.873)	.309	6_15	(.885)	.448
1_16	.799	.384	6_16	.840	.465
1_17	.719	.412	6_17	.685	.522
2_1	(.919)	.210	6_18	.730	.529
2_2	.715	.400	6_19	(.880)	.474
2_3	.458	.513	6_20	.719	.447
2_4	.638	.447	6_21	.662	.481
2_5	.627	.489	6_22	.825	.452
2_6	.784	.391	6_23	.754	.495
2_7	.773	.409	6_24	.747	.498
2_8	.557	.460	6_25	.744	.507
2_9	(.953)	.176	6_26	.427	.608
3_1	(.990)	.132	7_1	(.951)	.334
3_2	.680	.279	7_2	(.948)	.230
3_3	(.898)	.332	7_3	.760	.450
3_4	(.891)	.329	***7_4***	***(.167)***	***(.082)***
3_5	.815	.275	7_5	.516	.600
3_6	.612	.343	7_6	.751	.510
3_7	.715	.281	7_7	.682	.526
3_8	(.984)	.129	7_8	.582	.560
4_1	(.932)	.175	7_9	(.084)	(.155)
4_2	(.955)	.175	7_10	.437	.615
4_3	(.972)	.170	7_11	.439	.421
5_1	.789	.477	7_12	.821	.571
5_2	(.944)	.207	7_13	.712	.607
5_3	(.944)	.180	7_14	.652	.581
5_4	***.275***	***(-.049)***	7_15	.747	.586
5_5	***.202***	***(-.028)***	***7_16***	***(.028)***	***(-.033)***
5_6	(.909)	.262	7_17	(.051)	.152
5_7	(.855)	.305	7_18	.676	.643
5_8	.787	.493	7_19	.688	.640
5_9	.785	.367	7_20	.667	.651
5_10	.769	.487	7_21	.656	.643
5_11	.639	.483	7_22	.626	.620
5_12	.641	.488	7_23	.638	.595
5_13	.765	.541	7_24	.681	.647
5_14	.684	.539	7_25	.543	.628
5_15	.745	.532	7_26	.601	.600
5_16	(.950)	.163	7_27	.683	.536

TABLE 3. SUMMARY OF RESULTS FROM EFA USING A PRINCIPAL FACTOR ANALYSIS WITH A PROMAX ROTATION AND INTERNAL CONSISTENCY RELIABILITY (N = 1093).

Factor 1 (13)	Factor 2 (9)	Factor 3 (6)	Factor 4 (9)	Factor 5 (21)	Factor 6 (24)
1_1	2_1	3_1	5_1	6_2	7_1
1_2	2_2	3_3	5_8	6_6	7_2
1_3	2_3	3_4	5_9	6_8	7_3
1_4	2_4	3_5	5_10	6_9	7_5
1_6	2_5	3_6	5_11	6_10	7_6
1_8	2_6	3_7	5_12	6_11	7_7
1_9	2_7		5_13	6_12	7_8
1_10	2_8		5_14	6_13	7_9
1_11	2_9		5_15	6_14	7_10
1_14				6_15	7_11
1_15				6_16	7_12
1_16				6_17	7_13
1_17				6_18	7_14
				6_19	7_15
				6_20	7_18
				6_21	7_19
				6_22	7_20
				6_23	7_21
				6_24	7_22
				6_25	7_23
				6_26	7_24
					7_25
					7_26
					7_27
Inter-factor correlations					
Factor 1	.280	.118	.254	.302	.204
Factor 2		.422	.438	.396	.435
Factor 3			.319	.280	.352
Factor 4				.423	.366
Factor 5					.381
Reliability estimate (95% CI)					
.91 (.90, .92)	.89 (.88, .90)	.76 (.74, .78)	.94 (.93, .95)	.94 (.93, .94)	.96 (.96, .96)

The three game actions in stage 4 did not load high together on a single factor; they also failed to load with any other stage, and hence disappeared in the final factor solution. It is also noted that the six factors had correlations ranging from .12 to .45 (see middle panel of Table 3), indicating weak to moderate linear associations. This further supports a multi-dimensional structure as being measured by BTSG.

The bottom panel of Table 3 displays the internal consistency reliability estimates (together with their 95% CIs) for each factor extracted. It is clear from the table that the remaining stages (1, 2, 3, 5, 6 and 7) had internal consistency reliability estimates ranging from .76 (factor 3, which involved the smallest number of actions) to .96 (factor 6, which involved the largest number of actions). In effect, most of the estimates

were .89 or higher, representing good to excellent reliability. This further resulted in an overall internal consistency reliability estimate of .96 for the total 82 game actions after EFA.

VI. CONCLUSIONS

In this study, serious games analytics collected *in situ* BTSG served as the empirical data for analysis. We evaluated the following psychometric properties, including: (a) action analysis (using difficulty and discrimination measures), (b) construct validity (through an EFA with principal factor analysis and a Promax rotation), and (c) internal-consistency reliability (as Cronbach's alpha). After removing game actions that were not very discriminating or did not load on a single factor, the resulting game (with 82 game actions designed to measure objectives in six of the seven stages) was both internally reliable and valid in measuring six remotely related latent subconstructs.

Notwithstanding that data analysis indicated BTSG to be a viable assessment tool, not all decision makers in management and administration are psychometricians who interpret research reports with aplomb. What do the figures and numbers mean to them in terms of actionable insights and policies? Who among the nurses have met the competency requirements? Are there nurses who failed the competency assessment and should receive (re)training, or be placed on employment probation?

In the following section, we offer two options for analytics visualizations to provide actionable insights for decision makers. Chief Learning Officers of healthcare institutes can use these visualizations to promote successful trainees, recommend (re)training and assessments for those who need them, and even as feedback to revise serious games towards future studies.

A. Creating Competency Profiles of Trainees

For the six factors extracted using EFA, we recommend a radar chart that is capable of comparing several key indicators simultaneously for dashboard use. The radar chart can serve as a "competency profile" for the nurses, be it as an individual (singular) profile, a group profile for comparisons, or a summary of two or more individuals. Based on the game actions that loaded on each factor, the extracted factors can be named as follows:

- *Factor 1: Collect Blood* – collect blood from patient, fill out groups, and cross match form.
- *Factor 2: Order Blood* – verify details on groups, cross match form, and order correct blood from Blood Bank.
- *Factor 3: Blood Prepping* – chill blood box, and prepare blood bag for transportation.
- *Factor 4: Verify Match* – verify match between patient and blood.
- *Factor 5: Start Transfusion* – prep the machine, and start blood transfusion process.
- *Factor 6: Monitor Transfusion* – monitor transfusion to the end, and stop the machine.

As an example, we showed two nurses' competency profiles using their average raw scores for each factor in Figure 1. The figure easily allows visualizations of whom among the trainees (Trainee #1 or Trainee #2) performed better on which factors. It is obvious that Trainee #1 would require more training in blood transfusion as compared to Trainee #2. The competency profile for Trainee #1 also indicated more training particularly in Blood Prepping (factor 3: lowest score). In comparison, Trainee #2 performed generally well in all extracted factors, except in Blood Collection (factor 1). A closer analysis of the data showed Trainee #2 to have skipped over many game actions in stage 1, which explained the poor scores. In this case, Trainee #2 can be tasked with repeating the training and assessment for stage 1, without being subjected to a full assessment. This will not only save time and cost for the management, but also reduce stress for the trainee.

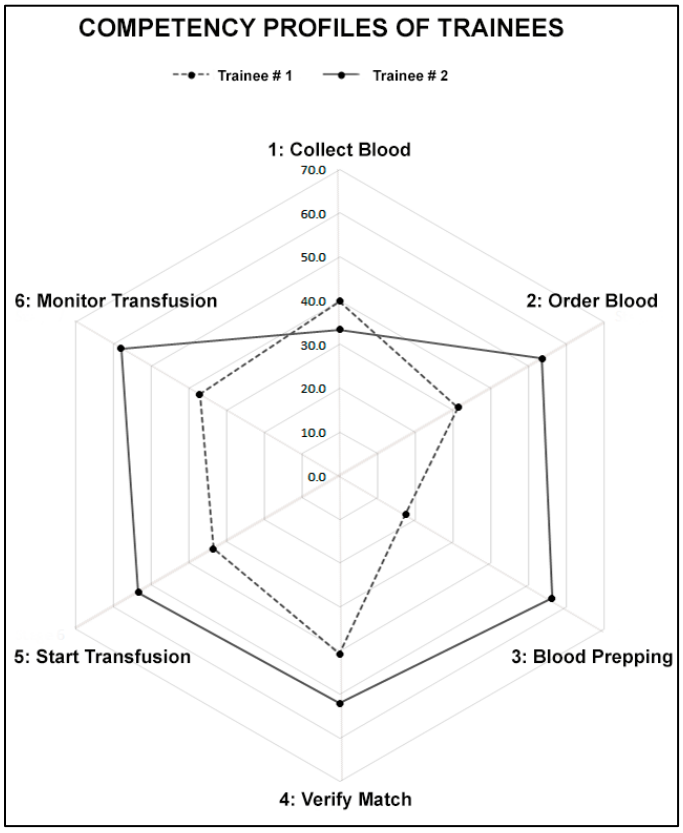

Fig. 1. Competency Profiles of Trainees (Two shown here for comparison)

Because three game actions from stage 4 failed to load well on any extracted factors, they were removed in the EFA process, which resulted in stage 4 disappearing altogether. It so happens that the game actions from each stage roughly loaded onto each extracted factor, where: stage 1 actions → factor 1, stage 2 actions → factor 2, stage 3 actions → factor 3, stage 5 actions → factor 4, stage 6 actions → factor 5, and stage 7 actions → factor 6. We must stress that game actions can load multi-dimensionally onto more than one factor, and that the apparent correspondence between factors and stages is a mere coincidence for this study. Researchers should not expect a similar corresponding effect for other serious games.

From an instructional design perspective, stage 4 of the current version of BTSG contained too few game actions (just three) to make any statistical impact from measurement considerations. Our recommendation for the game developer

would be to either redesign stage 4 to comprise more 'game actions', or remove stage 4 altogether. Similarly, factor 3 has the least reliability estimate (.76, 95%CI: .74, .78) (see Table 3), which could be due to the small number of game actions (just six) that loaded onto the factor. Adding more game actions to stage 3 could help improve the reliability estimate of this stage.

B. Group Competency Profile: Boxplot and Outliers

We recommend using a boxplot to visualize players' average raw scores by factor (Figure 2). Because a boxplot reveals not only the location, scale, and symmetry within a dataset, but also extreme scores or potential outliers, it is considered to be very useful for visualizations of competency assessments. For example, Fig. 2 shows that a great majority of players performed very well on the six factors, with median scores >.75, which speaks well of the high competencies of the participants of the department. Further, we note that players performed worse on factors 2 and 6 (whose first quartiles (Q1) are <.5), as compared with factors 1, 3, 4, and 5 (whose Q1 values are all >.6).

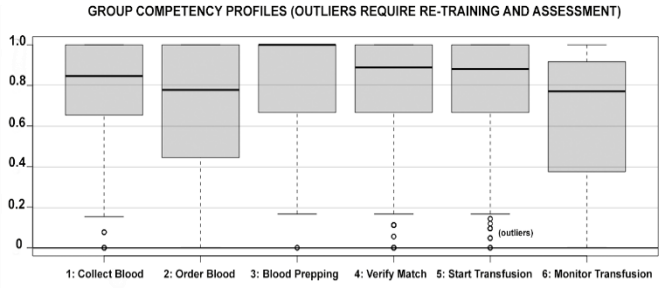

Fig. 2. Boxplot of average raw score.

All factors, except for 2 and 6, show a number of outliers – denoted as small circles on the bottom of each boxplot (see Figure 2). In a boxplot, those who have performed lower than one step (1.5 x interquartile range) below Q1 are outliers. In other words, these (outlier) nurses under-performed by some margin as compared to the rest of the nurses in that competency (i.e., factor).

Fig. 2 also shows factors 2 and 6 with relatively larger variations in the middle of the distributions of average raw scores. This means that for the nurses, the greatest variation in the middle of the competency distributions can be found in factors 2 and 6 (i.e., Order Blood, and Monitor Transfusion). While over 50% of the nurses have average competencies in Order Blood or Monitor Transfusion (>.6), there are those who are very low in these competencies also. The relatively larger variation lowered Q1 in the distributions of these two factors, which in turn explains why there are no outliers for factors 2 and 6 (as the minimum possible for each stage's mean score = 0).

On the other hand, the median of factor 3 topped out at the maximum value (1). This means at least 50% of the nurses responded correctly on all game actions related to this factor. It can mean that, either the Blood Prepping competency (factor 3) was relatively easier than the other competencies, or that the majority of the nurses have relatively higher competency in Blood Prepping.

The number of outliers who under-performed in each factor of BTSG is provided in Table 4. When summed, the total number of outliers for the six factors is 313 cases; however, analysis reveals only 263 unique IDs. This indicates that there are 'repeated outliers' – meaning, some trainees under-performed in more than one factor. A breakdown of the number of trainees (outliers) who under-performed in multiple factors is shown in Table 5.

TABLE 4. NUMBER OF OUTLIERS FOR EACH EXTRACTED FACTOR.

1: Collect Blood	2: Order Blood	3: Blood Prepping	4: Verify Match	5 Start Transfusion	6: Monitor Transfusion
60	0	9	186	58	0

* There are a total of 313 outlier cases, but only 263 unique IDs.

TABLE 5. NUMBER OF OUTLIERS FOR EACH COMBINATION OF FACTORS.

Factors	(1,3)	(1,4)	(1,5)	(3,4)	(3,5)	(4,5)	(1,3,4)	(1,3,5)	(3,4,5)	(1,3,4,5)
Outliers	2	11	7	4	3	30	1	1	3	1

The highest number (30) of under-performing trainees was found in the factor-combination (4, 5), which suggested this group of trainees were unfamiliar with the processes leading up to getting the patient ready to receive blood transfusion (factors 4, 5: Verify Match, and Start Transfusion). From Table 4, we can see that factor 4 has the most outliers (186); we would recommend they be sent for re-training in this competency (4: Verify Match). It is further noted that one particular nurse under-performed in multiple factors (1, 3, 4, 5); but given that factors 2 and 6 have a larger variability, it is very likely this person has, in fact, under-performed in all factors! A drill down analysis revealed this was indeed the situation: the person had 0 average scores on all factors. If SGH really adopts BTSG to be an official assessment tool, this one nurse should probably be placed under probation right away (or be dismissed), as such (gross) underperformance would not be acceptable for a registered nurse. As for the rest of the outliers, management may choose to place them under probation for observation, or re-training and assessment, if so desired.

C. Summary

In conclusion, BTSG is a viable assessment tool for BTA competency. Certainly, further validation evidence for the BTSG needs to be gathered by seeking other validation evidence or using other hospitals in Singapore. While this paper focused only on nurses as participants, the demonstrated measurement practices and visualization strategies can be generalized to other healthcare workers who also require the knowledge and techniques. Visualizations of analytics are decidedly helpful in not only filling the analytics dashboard, but also in providing insights on actionable items for management and administration to make better decisions when it comes to training and assessment. We believe the clear evidence provided by studies like this will go a long way in providing management and administration the data they need to convince shareholders to invest in serious games for future training.

REFERENCES

[1] P. M. Kato and S. de Klerk, "Serious games for assessment: Welcome to the jungle," J. Appl. Test. Technol., vol. 18, no. S1, pp. 1–6, 2017.

[2] M. Bauer et al., "Why video games can be a good fit for formative assessment," J. Appl. Test. Technol., vol. 18, no. S1, pp. 19–31, 2017.

[3] Singapore Nursing Board, Core Competencies and Generic Skills of Registered Nurses, no. April. Singapore, 2018. [Online]. Available: https://www.healthprofessionals.gov.sg/docs/librariesprovider4/publications/core-competencies-generic-skills-of-rn_snb_april-2018.pdf

[4] A. Karami, J. Farokhzadian, and G. Foroughameri, "Nurses' professional competency and organizational commitment: Is it important for human resource management?," PLoS One, vol. 12, no. 11, p. e0187863, Nov. 2017, doi: 10.1371/journal.pone.0187863.

[5] M. Fukada, "Nursing competency: Definition, structure and development," Yonago Acta Med., vol. 61, no. 1, pp. 001–007, 2018, doi: 10.33160/yam.2018.03.001.

[6] A. A. Bediako, R. Ofosu-Poku, and A. A. Druye, "Safe blood transfusion practices among nurses in a major referral center in Ghana," Adv. Hematol., vol. 2021, pp. 1–13, Mar. 2021, doi: 10.1155/2021/6739329.

[7] E. Lancaster, E. Rhodus, M. Duke, and A. Harris, "Blood transfusion errors within a health system: A review of root cause analyses," Patient Saf., pp. 78–91, Jun. 2021, doi: 10.33940/med/2021.6.6.

[8] Health Science Authority, Clinical Blood Transfusion. 2011.

[9] H. Forsman et al., "Clusters of competence: Relationship between self-reported professional competence and achievement on a national examination among graduating nursing students," J. Adv. Nurs., vol. 76, no. 1, pp. 199–208, Jan. 2020, doi: 10.1111/jan.14222.

[10] C. S. Loh, Y. Sheng, and D. Ifenthaler, Eds., Serious Games Analytics. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-05834-4.

[11] C. S. Loh and Y. Sheng, "Measuring expert performance for serious games analytics: From data to insights," in Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement, C. S. Loh, Y. Sheng, and D. Ifenthaler, Eds., Switzerland: Springer International Publishing, 2015, pp. 101–134. doi: 10.1007/978-3-319-05834-4_5.

[12] C. S. Loh, Y. Sheng, and D. Ifenthaler, "Serious games analytics: Theoretical framework," in Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement, C. S. Loh, Y. Sheng, and D. Ifenthaler, Eds., Switzerland: Springer International Publishing, 2015, pp. 3–29. doi: 10.1007/978-3-319-05834-4_1.

[13] D. P. Thangavelu, A. J. Q. Tan, R. Cant, W. L. Chua, and S. Y. Liaw, "Digital serious games in developing nursing clinical competence: A systematic review and meta-analysis," Nurse Educ. Today, vol. 113, p. 105357, Jun. 2022, doi: 10.1016/j.nedt.2022.105357.

[14] M. Graafland et al., "How to systematically assess serious games applied to health care," JMIR Serious Games, vol. 2, no. 2, p. e11, Nov. 2014, doi: 10.2196/games.3825.

[15] B. Sawyer and D. Rejeski, "Serious games: Improving public policy through game-based learning and simulation," Washington, DC., 2002. [Online]. Available: https://www.wilsoncenter.org/sites/default/files/media/documents/publication/ACF3F.pdf

[16] L. Pront, A. Müller, A. Koschade, and A. Hutton, "Gaming in nursing education: A literature review," Nurs. Educ. Perspect., vol. 39, no. 1, pp. 23–28, Jan. 2018, doi: 10.1097/01.NEP.0000000000000251.

[17] H. M. Johnsen, M. Fossum, P. Vivekananda-Schmidt, A. Fruhling, and Å. Slettebø, "Developing a serious game for nurse education," J. Gerontol. Nurs., vol. 44, no. 1, pp. 15–19, Jan. 2018, doi: 10.3928/00989134-20171213-05.

[18] A. Min, H. Min, and S. Kim, "Effectiveness of serious games in nurse education: A systematic review," Nurse Educ. Today, vol. 108, p. 105178, Jan. 2022, doi: 10.1016/j.nedt.2021.105178.

[19] A. J. Q. Tan, C. C. S. Lau, and S. Y. Liaw, "Serious games in nursing education: An integrative review," in 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), IEEE, Sep. 2017, pp. 187–188. doi: 10.1109/VS-GAMES.2017.8056599.

[20] S. Göbel, O. Hugo, M. Kickmeier-Rust, and S. Egenfeldt-Nielsen, "Serious games — Economic and legal issues," in Serious Games, Cham: Springer International Publishing, 2016, pp. 303–318. doi: 10.1007/978-3-319-40612-1_11.

[21] E. Hariton and J. J. Locascio, "Randomised controlled trials - the gold standard for effectiveness research," BJOG An Int. J. Obstet. Gynaecol., vol. 125, no. 13, pp. 1716–1716, Dec. 2018, doi: 10.1111/1471-0528.15199.

[22] R. E. Clark, "Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses," Educ. Commun. Technol., vol. 33, no. 4, pp. 249–262, 1985, doi: 10.1007/BF02769362.

[23] R. E. Clark, "Confounding in Educational Computing Research," J. Educ. Comput. Res., vol. 1, no. 2, pp. 137–148, 1985, doi: 10.2190/hc3l-g6yd-bak9-eqb5.

[24] R. E. Clark, "Reconsidering research on learning from media," Rev. Educ. Res., vol. 53, no. 4, pp. 445–459, 1983, doi: 10.3102/00346543053004445.

[25] R. E. Clark, "Dangers in the evaluation of instructional media," Acad. Med., vol. 67, no. 12, pp. 819–820, 1992, [Online]. Available: http://journals.lww.com/academicmedicine/Fulltext/1992/12000/Dangers_in_the_evaluation_of_instructional_media_.4.aspx

[26] M. Graafland, J. M. Schraagen, and M. P. Schijven, "Systematic review of serious games for medical education and surgical skills training.," Br. J. Surg., vol. 99, no. 10, pp. 1322–1330, Oct. 2012, doi: 10.1002/bjs.8819.

[27] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," Nature, vol. 567, no. 7748, pp. 305–307, 2019, [Online]. Available: https://www.nature.com/articles/d41586-019-00857-9

[28] N. B. Hastings and M. W. Tracey, "Does media affect learning: where are we now?," TechTrends, vol. 49, no. 2, pp. 28–30, 2005, doi: 10.1007/bf02773968.

[29] T. Zhou and C. S. Loh, "The effects of fully and partially in-game guidance on players' declarative and procedural knowledge with a disaster preparedness serious game," Int. J. Gaming Comput. Simulations, vol. 12, no. 4, pp. 23–37, Oct. 2020, doi: 10.4018/IJGCMS.2020100102.

[30] S. de Klerk and P. M. Kato, "The future value of serious games for assessment: Where do we go now?," J. Appl. Test. Technol., vol. 18, no. S1, pp. 32–37, 2017.

[31] L. W. T. Schuwirth and C. P. M. Van Der Vleuten, "Current assessment in medical education: Programmatic assessment," J. Appl. Test. Technol., vol. 20, no. S2, pp. 2–10, 2019.

[32] A. Blanié, M.-A. Amorim, A. Meffert, C. Perrot, L. Dondelli, and D. Benhamou, "Assessing validity evidence for a serious game dedicated to patient clinical deterioration and communication," Adv. Simul., vol. 5, no. 4, 2020, doi: 10.1186/s41077-020-00123-3.

[33] J. M. Gerard et al., "Validity evidence for a serious game to assess performance on critical pediatric emergency medicine scenarios," Simul. Healthc. J. Soc. Simul. Healthc., vol. 13, no. 3, pp. 168–180, 2018, doi: 10.1097/SIH.0000000000000283.

[34] T. J. Olgers, J. M. van Os, H. R. Bouma, and J. C. ter Maaten, "The validation of a serious game for teaching ultrasound skills," Ultrasound J., vol. 14, no. 29, 2022.

[35] F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Boston, MA: Addison-Wesley, 1968.

[36] AERA, APA, and NCME, The standards for educational and psychological testing. Washington, DC: AERA Publications Sales, 1999.

[37] D. A. Cook and T. J. Beckman, "Current concepts in validity and reliability for psychometric instruments: theory and application," Am. J. Med., vol. 119, no. 2, pp. 166.e7–16, 2006, doi: 10.1016/j.amjmed.2005.10.036.