Enhancing the Assessment of Serious Games using Fuzzy AHP

Omari Kamal

ABSTRACT

In recent years, Serious Games have emerged as an innovative pedagogical strategy for enhancing learning in various fields of knowledge. However, in order to effectively utilize Serious Games, it is crucial to systematically evaluate them and obtain robust evidence of their impacts. To achieve this, integrating Fuzzy Multi-Criteria Decision Making (FMCDM) methods can be beneficial for evaluating Serious Games, as they allow for the weighting of evaluation criteria based on the specific context of use. FMCDM methods have the ability to account for the imprecision and uncertainty of human judgments. The objective of this paper is to propose an evaluation system for Serious Games based on four dimensions that are validated and weighted using the fuzzy Analytic Hierarchy Process (AHP) method, tailored to the context of use. The proposed Serious Games evaluation system was tested on a Serious Game validated by a pedagogical committee, and the results obtained demonstrate the quality and relevance of the proposed evaluation system for Serious Games.

Keywords: FMCDM, Fuzzy AHP, Pedagogical Strategy, Serious Game.

O. Kamal*
Department of Mathematics & Computer Science, Faculty of Sciences Ben M'sik, University Hassan II Casablanca, Morocco.
(e-mail: kamal.omari2013@gmail.com)

*Corresponding Author

I. INTRODUCTION

In recent years, Serious Games (SGs) have emerged as an innovative pedagogical strategy for more effective and motivating learning [1]. However, in order to utilize SGs optimally, it is essential to evaluate them and obtain solid evidence of their impacts [2]. Therefore, a suitable evaluation system is necessary [3] to assess the quality of their design [4], diagnose their usage [5], and assess their educational impact [6]. Without proper evaluation, there is no proof that the intended purpose of a SG has been achieved.

Hence, the objective of this paper is to propose a SGs evaluation system that is designed around four essential dimensions that a SG must fulfill to achieve its intended task. These dimensions are weighted and validated using the Fuzzy Analytic Hierarchy Process (AHP), a decision-making technique that accounts for imprecision and uncertainty in judgments, to ensure internal coherence in the evaluator's judgments based on the specific context of SG usage.

II. STATE OF THE ART

The establishment of Serious Games (SGs) as a pedagogical solution may not be feasible without proper evaluation. Therefore, many researchers have proposed frameworks and methodologies for evaluating SGs, focusing on their potential, quality, or effectiveness. For instance, [7] developed a four-dimensional model of SG evaluation that includes context, learner specification, representation mode, and pedagogical considerations.

Reference [8] proposed an evaluation framework based on a questionnaire to stimulate critical discourse on the potential of SGs, emphasizing the cohesion between essential design elements and the SG objective.

In addition, [9] proposed a measurement scale to assess the quality of SGs before implementation in training, measuring the pleasure during the game and perceived learning. Furthermore, [10] a four-section evaluation and analysis grid was proposed to assess the quality of SGs intended for educational use, covering identification of the SG, educational specifications, playful specifications, and technical specifications.

Various methods and techniques are employed to evaluate the effectiveness of SGs. Reference [11] presented an evaluation framework that includes history, mechanisms, usability, knowledge, motivation, and satisfaction as criteria, covering the first two levels of Kirkpatrick's evaluation model [12]. Reference [13] proposed a methodology for non-disruptive monitoring in SGs to measure learning outcomes and systematically evaluate their effectiveness.

Based on this literature review, it is evident that several criteria have been addressed in different frameworks and methodologies for SG evaluation. However, there is a lack of an evaluation system specifically designed to assess SGs as the outcome of tool development projects intended for training contexts. Therefore, in this paper, we propose a new evaluation system.

III. DIMENSIONS OF SG EVALUATION

The proposed SG evaluation system in this paper is based on four dimensions that are considered essential for evaluating a serious game (SG), and these dimensions will be measured according to well-defined criteria. The four dimensions are:

Pedagogical Dimension (PD): This dimension focuses on

evaluating the educational content provided by the SG. The measurement criteria for this dimension include:

- Targeted Skills (Ts): Evaluating whether the SG effectively targets the desired learning outcomes and skills.
- Pedagogical Considerations (Pc): Evaluating whether the SG incorporates sound pedagogical principles and instructional design strategies.
- Learning Results (Lr): Evaluating the effectiveness of the SG in facilitating actual learning outcomes.
- Error Management (Em): Evaluating how the SG handles errors or mistakes made by the learners and provides feedback for learning improvement.

Technological Dimension (TD): This dimension focuses on evaluating the technological aspects of the SG. The measurement criteria for this dimension include:

- Game Design (Gd): Evaluating the overall design of the SG, including its visual aesthetics, sound, and overall game mechanics.
- Performance (P): Evaluating the technical performance of the SG, such as its loading times, responsiveness, and stability.
- User Interface (Ui): Evaluating the user interface design of the SG, including its ease of use, navigation, and accessibility.
- Usability (U): Evaluating the overall usability of the SG, including its user-friendliness, ease of learning, and user satisfaction.

Ludic Dimension (LD): This dimension focuses on evaluating the interactive components of the SG that contribute to the motivation and engagement of the learners/players. The measurement criteria for this dimension include:

- Challenge (C): Evaluating the level of challenge and difficulty of the SG, which can influence the motivation and engagement of the learners/players.
- Fun (F): Evaluating the enjoyment and entertainment value of the SG, which can contribute to the overall motivation and engagement of the learners/players.
- Gameplay (G): Evaluating the quality and effectiveness of the gameplay mechanics, including the interactivity, game mechanics, and game progression.
- Immersion (I): Evaluating the degree of immersion and engagement experienced by the learners/players within the SG environment.

Behavioural Dimension (BD): This dimension focuses on evaluating the involvement and attitude of the learners/players towards the SG. The measurement criteria for this dimension include:

- Motivation (M): Evaluating the motivation levels of the learners/players throughout the SG experience, including their intrinsic motivation, extrinsic motivation, and overall engagement.
- Engagement (E): Evaluating the level of engagement and involvement of the learners/players in the SG, including their active participation, attention, and interest.
- User Experience (Ue): Evaluating the overall experience of the learners/players with the SG, including their satisfaction, feedback, and perception of the SG.

This SGs evaluation system is modelled by the set of evaluation dimensions A and the measurement criteria A_i:

$A = \{PD, TD, LD, BD\}$
$A_{PD} = \{Ts, Pc, Lr, Em\}$,
$A_{TD} = \{Gd, P, Ui, U\}$,
$A_{LD} = \{C, F, G, I\}$,
$A_{BD} = \{M, E, Ue\}$.

By evaluating the SG based on these four dimensions and their respective criteria, the proposed SG evaluation system aims to provide a comprehensive and holistic assessment of the SG's effectiveness and quality as a pedagogical tool in a training context.

IV. CRITERIA WEIGHTING METHOD

The SG evaluation system proposed in this paper acknowledges the complexity of evaluating SGs with multiple criteria and the need to assign appropriate weights to these criteria. To address this, a multi-criteria decision-making (MCDM) method is employed, such as the Decision Making Trial and Evaluation Laboratory (DEMATEL) method [14], Stepwise Weight Assessment Ratio Analysis (SWARA) method [15], or Analytical Hierarchy Process (AHP) method [16]. These methods allow for a systematic approach to assigning weights to criteria based on their relative importance.

In addition, the fuzzy logic proposed by [17] is integrated into the evaluation system to handle the imprecision and ambiguity of data often encountered in human decision-making. Specifically, the fuzzy AHP method described by [18] is utilized, which employs linguistic variables represented by triangular fuzzy numbers.

A triangular fuzzy number is denoted by a triplet (l, m, u), where l ≤ m ≤ u. The values of l and u represent the lower and upper bounds of the support, respectively, and m represents the modal value. This approach allows for a more flexible and nuanced representation of the uncertainty and vagueness associated with evaluating the criteria in the SG evaluation system.

By incorporating the fuzzy AHP method, the proposed SG evaluation system aims to account for the inherent imprecision and ambiguity in evaluating SGs and to provide a more robust and comprehensive approach to assigning weights to criteria, considering their relative importance in the evaluation process.

Its membership function is defined as follows:

$$U_{\tilde{M}} = \begin{cases} 0, & x < l, \\ \dfrac{(x - l)}{(m - l)}, & l < x \leq m, \\ \dfrac{(u - x)}{(u - m)}, & m \leq x \leq u, \\ 0, & x > u \end{cases}$$

V. PROCESS OF SG EVALUATION

The proposed SG evaluation process involves the evaluator comparing the criteria with each other using linguistic values. This comparison is processed to obtain appropriate weightings for each criterion, based on the evaluator's choice,

while maintaining coherence among the weights. The coherence ratio, which should not exceed 0.10, is used as a measure of consistency in the weightings assigned by the evaluator.

Once the criteria are weighted, students are asked to play the selected SG. After playing the SG, students answer a pre-set multiple-choice questionnaire (MCQ) that is designed by the evaluator. Finally, the evaluator analyzes the results from the MCQ to generate evaluation outcomes (as shown in Fig. 1).
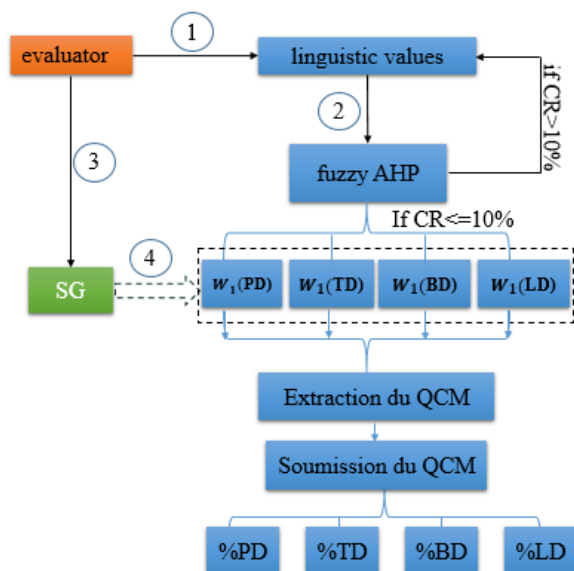

Fig. 1. Evaluation process for SGs.

1) The evaluator makes their choice by pairwise comparisons of the evaluation aspects {PD, TD, LD, BD}.
2) The linguistic values chosen by the evaluator are processed using the fuzzy AHP method.
3) The evaluator determines the Serious Game (SG) to be evaluated before placing it in a formative context.
4) The SG is evaluated based on the weighted aspects.

This process allows for a comprehensive evaluation of the SGs, taking into account the subjective judgments of the evaluator in assigning weights to criteria, as well as the feedback from students through the MCQ. The combination of evaluator judgments and student feedback provides a holistic assessment of the SGs in terms of pedagogical, technological, ludic, and behavioral dimensions, as well as their overall effectiveness in achieving the intended learning objectives.

VI. Application of The Sg Evaluation System

A. The process of Weighting Criteria

The process of weighting the criteria for SG evaluation using the fuzzy AHP method consists of seven steps:

Step 1: (pairwise comparison between criteria)

As an evaluator, we compared the dimensions of set A, by giving a language value for each comparison.

TABLE I: LINGUISTIC VALUE MAPPING - TRIANGULAR FUZZY NUMBER

Pairwise	Linguistic	Importance	Triangular

comparison	value	relative	fuzzy number $\tilde{P}_{ij}$
(PD, TD)	Moderate importance	$\tilde{3}$	[2, 3, 4]
(PD, BD)	High importance	$\tilde{5}$	[4, 5, 6]
(PD, LD)	Very high importance	$\tilde{7}$	[6, 7, 8]
(TD, BD)	Moderate importance	$\tilde{3}$	[2, 3, 4]
(TD, LD)	High importance	$\tilde{5}$	[4, 5, 6]
(BD, LD)	Moderate importance	$\tilde{3}$	[2, 3, 4]

As mentioned in Table I, we thought it wise to privilege PD over all other dimensions, as our context is purely formative. Moreover, since the target population is of university and scientific level, therefore, they are used to new information technologies, which pushed us to privilege TD over BD and LD. In addition, the SG is used in an educational activity, so we have favored BD over LD.

Then these values will be transformed into triangular fuzzy numbers $\tilde{P}_{ij}$.

Step 2: (Construction of the fuzzy judgment matrix $\widetilde{M}$)

We define the fuzzy judgment matrix $\widetilde{M}$ composed by the triangular fuzzy numbers $\tilde{P}_{ij}\{l_{ij}, m_{ij}, u_{ij}\}$, this matrix contains the aggregation of all triangular fuzzy numbers $\tilde{P}_{ij}$.

$$\widetilde{M} = \begin{pmatrix} (1,1,1) & \cdots & \tilde{P}_{ij} \\ \vdots & \ddots & \vdots \\ \tilde{P}_{ji} & \cdots & (1,1,1) \end{pmatrix} \qquad (1)$$

With $\tilde{P}_{ij}=\tilde{P}_{ji}^{-1}$ and $\tilde{P}_{ji}^{-1} = (\frac{1}{u_{ij}},\frac{1}{m_{ij}},\frac{1}{l_{ij}})$

Step 3: (Calculation of the consistency ratio)

During this stage, the evaluation system may recommend that we redo the initial comparisons, with the aim of maintaining a consistency ratio that must not exceed 0.10 according to [19], in order to validate the choices made.

This index is defined as a ratio between the coherence index of the evaluation matrix (CI) and the coherence index of a random matrix (RI).

$$CI = \frac{\lambda_{max}-n}{n-1} \ and \ CR = \frac{CI}{RI} \leq 0.10 \qquad (2)$$

With (λ_{max}) is the eigenvalue and (n) the number of criteria.

In our case, we have $\lambda_{max}= 4.154$, n = 4 and CI = 0.051. Therefore, CR = 0.570≤0.10.

This means that our choice of pairwise comparison is valid.

Step 4: (Calculation of the geometric mean)

The system calculates the fuzzy geometric mean for each criterion with the equation below:

$$\tilde{r}_i = \left(\prod_{j=1}^n \tilde{P}_{ij}\right)^{\frac{1}{n}} ; I = 1,2, \ldots\ldots, n \qquad (3)$$

Step 5: (Calculation of the fuzzy weights)

The fuzzy weight of each criterion (w_i) is calculated by multiplying each fuzzy geometric mean by the inverse of its vector, by the following equation.

$$\tilde{w}_i = \tilde{r}_i \otimes \left(\sum_{i=1}^{n} \tilde{r}_i\right)^{-1} \quad I = 1,2, \ldots, n \qquad (4)$$

Step 6: (defuzzification of the fuzzy weights)

The defuzzification of the fuzzy weights are done by applying the equation below.

$$M_i = \frac{lw_i + mw_i + uw_i}{3} \quad i=1,2,\ldots, n \qquad (5)$$

Step 7: (Normalization of the fuzzy weights)

The normalized fuzzy weights are obtained by using the equation below.

$$N_i = \frac{lw_i + mw_i + uw_i}{\sum_{i=1}^{n} M_i} \quad i = 1,2, \ldots, n \qquad (6)$$

After applying the above steps to all criteria, the following results are obtained:
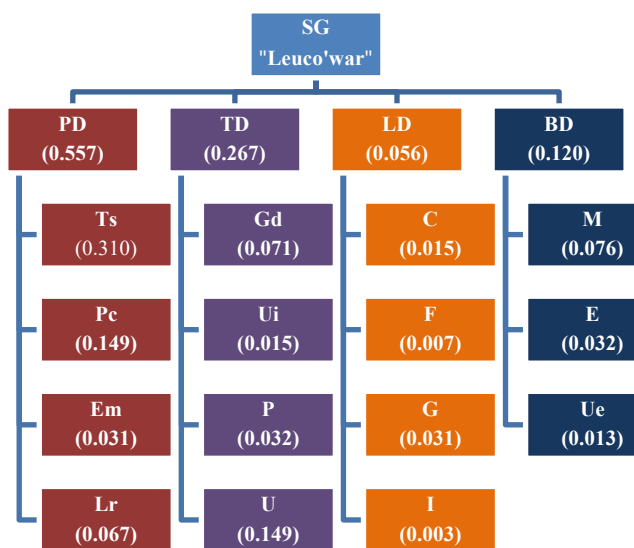

Fig. 2. The weightings of the SG evaluation system.

B. The SG Chosen

The pedagogical committee of the life sciences sector at Hassan II University of Casablanca has approved the use of the SG called "Leuco'war" to experiment with our SG evaluation system.

Philippe Cosentino created this SG «Leuco'war», on the theme of "interactions between different leucocytes".

In this SG, the student takes on the role of a nano-doctor tasked with assisting a patient's immune system in healing as quickly as possible. Throughout the game, the student will learn about macrophages, monocytes, mast cells, and B-lymphocytes.
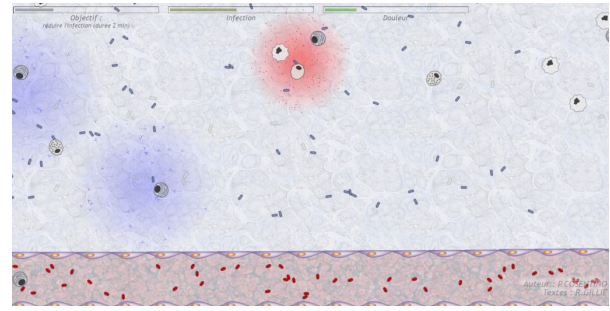

Fig. 3. The SG Leuco'war.

VII. Results and Discussion

After the students played the SG "Leuco'war", they were asked to complete the proposed multiple-choice questionnaire (MCQ). The calculated Cronbach's Alpha value for the MCQ was found to be 0.954, indicating high consistency among the questionnaire items. This value exceeds the minimum threshold of 0.70 [20], indicating strong internal consistency of the questionnaire, and suggesting that the MCQ is a reliable measure of the students' responses.

Fig. 4. Analysis results of the SG "Leuco'war.

The results of the MCQ analysis revealed that the SG "Leuco'war" is more suitable for use in an entertainment context rather than solely as a formative tool. The ludic (gameplay) dimension of the SG scored an average of 88.19%, which is higher than the pedagogical dimension. This suggests that students perceive the SG primarily as an entertaining activity rather than a purely educational one.

It is also noteworthy that in the design of the SG "Leuco'war", priority was given to the technological and behavioral dimensions over the pedagogical dimension. This indicates that the SG was primarily designed to incorporate technological features and behavioral aspects, such as interactions between different leucocytes, rather than focusing solely on pedagogical objectives. This may explain the higher emphasis on the ludic dimension and the lower score for the pedagogical dimension in the MCQ analysis results.

Overall, the findings suggest that "Leuco'war" may be better suited for use in an entertainment context, where the ludic dimension can be leveraged to engage students, rather than being used solely as a formative tool in an educational setting. Further analysis and evaluation may be needed to determine the optimal context for utilizing this SG in the future.

VIII. CONCLUSION

In this article, we have presented our Serious Games (SGs) evaluation system, which is designed around four dimensions: pedagogical, technological, behavioral, and ludic dimensions. To weight and validate these dimensions, we utilized the fuzzy multi-criteria decision-making method called "Fuzzy AHP" under a program developed in Matlab.

The results obtained from the experimentation of our SGs evaluation system revealed the benefits of using the Fuzzy AHP method in the criteria weighting process. It allowed for a flexible and adaptable SGs evaluation system that can be tailored to the specific context of use. The use of Fuzzy AHP ensured that the evaluation system is capable of accommodating the complex and subjective nature of SGs, making it suitable for diverse applications.

In our future work, we propose extending the application of our SGs evaluation system to different types of Serious Games. This would allow for further validation and refinement of the system, as well as its applicability to various educational and training domains.

REFERENCES

[1] Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education.* 2012; 59(2): 661-686.

[2] Petri G, Giani, Gresse von Wangenheim C. How to Evaluate Educational Games: a Systematic Literature Review. *Journal of Universal Computer Science.* 2016; 22: 992.

[3] Liu S, Ding W. An Approach to Evaluation Component Design in Building Serious Game. Learning by Playing: Game-based Education System Design and Development. *Edutainment*. 2009; 5670: 141-148.

[4] Di Loreto I, Gouaich A. An Early Evaluation Method for Social Presence in Serious Game. Proceedings of the Second International Conference on Computer Supported Education, eds. J.A. Moinhos Cordeiro, B. Shishkov, A. Verbraeck, & M. Helfert, Valencia, Spain, 2010.

[5] Loh CS. Information Trails: In-Process Assessment of Game-Based Learning. In: Ifenthaler, D., Eseryel, D., Ge, X. (eds) Assessment in Game-Based Learning. Springer, New York, 2012.

[6] Papastergiou M. Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation*. Computers & Education.* 2009; 52 (1): 1-12.

[7] De Freitas S, Oliver M. How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education.* 2006; 46: 249-264.

[8] Mitgutsch K, Alvarado N. Purposeful by design? : A serious game design assessment framework. Foundations of Digital Games, pp. 121-128. New York, USA, 2012.

[9] Ak, O. A Game Scale to Evaluate Educational Computer Games, Procedia-Social and Behavioral Sciences, eds Gülsün A. ,Fezile O, Sezer K; Deniz Ö, Barcelona, 2012.

[10] El Borji Y, Khaldi M. Comparative Study to Develop a Tool for the Quality Assessment of Serious Games Intended to be Used in Education. *International Journal of Emerging Technologies in Learning (iJET).* 2014; 9(9): 50-55.

[11] Carvalho C. Is game-based learning suitable for engineering education?. Proceeding of the Global Engineering Education Conference, Marrakech, Morocco: IEEE, 2012.

[12] Kirkpatrick DL, Kirkpatrick JD. Evaluating training programs: the four levels, 3rd ed. San Francisco, CA: Berrett-Koehler Publishers; 2006.

[13] Serrano-Laguna Á, Manero B, Freire M, Fernández-Manjón B. A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes. *Multimedia Tools and Applications*. 2018; 77(2): 2849-2871.

[14] Gabus A, Fontela E. World Problems an Invitation to Further Thought within the Framework of DEMATEL; Battelle Geneva Research Centre: Geneva, Switzerland, 1972.

[15] Keršuliene V, Zavadskas EK, Turskis Z. Selection of rational dispute resolution method by applying new step-wise weight assessment ratio analysis (SWARA). *J. Bus. Econ. Manag.* 2010; 11(2): 243-258.

[16] Saaty TL. Analytic Hierarchy Process. New York: McGraw-Hill; 1980.

[17] Zadeh LA. Fuzzy sets. *Information and Control.* 1965; 8(3): 338-353.

[18] Buckley JJ. Fuzzy hierarchical analysis. *Fuzzy Sets and Systems.* 1985; 17(3): 233-247.

[19] Saaty TL. Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World. 3rd ed. Pittsburgh: RWS Publications; 2012.

[20] DeVellis RF. Scale development: theory and applications. 2nd ed. Thousand Oaks, CA: Sage Publications; 2003.