A Metric Suite for Evaluating Interactive Scenarios in Video Games: An Empirical Validation

Maria-Eleni Paschali, Apostolos Ampatzoglou, Remi Escourrou, Alexander Chatzigeorgiou, Ioannis Stamelos

Computer Science Department, Aristotle University of Thessaloniki, Greece
Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
Department of Computer Science and Electronics, École Nationale Supérieure des MINES, France
mpaschali@csd.auth.gr, apostolos.ampatzoglou@gmail.com, remi.orious@orange.fr, achat@uom.gr, stamelos@csd.auth.gr

ABSTRACT

Game development is one of the fastest growing industries. Since games' success is mostly related to users' enjoyment, one of the cornerstones of their quality assessment is the evaluation from the user perspective. According to literature, game scenario constitutes a key-factor that leads to users' enjoyment. Despite their importance, scenarios are currently evaluated through heuristics in a subjective way. The aim of this paper is to develop an objective model (i.e., a set of quality attributes and metrics) for evaluating game scenarios with respect to users' satisfaction. The proposed model can be applied to flow charts and character models (i.e., common game scenario representation mechanisms). To achieve this goal, we: (a) gathered game scenario characteristics that are related to users' satisfaction, (b) proposed several metrics for quantifying these characteristics, and (c) performed a case study on three interactive scenarios to evaluate the model. As a result, we developed a three-level model: the first level includes high-level characteristics (e.g., interestingness), which are specified in the second level; the third level maps graph-based metrics to the attributes of the second level. The results of the empirical validation suggest that in the majority of the cases, the proposed metrics were strongly correlated with the perceived opinion of evaluators. The results can be useful to both researchers and practitioners, in the form of early quality assessment instruments (regarding practitioners) and future research directions (regarding researchers).

CCS CONCEPTS

Software and its engineering→ Software creation and management→ Software verification and validation→ Empirical software validation

KEYWORDS

Computer games, interactive scenarios, evaluation, metrics

SAC '20, 30 March – 3 April 2020, Brno, Czech Republic

ACM ISBN
https://doi.org/

ACM reference format

M. E. Paschali, A. Ampatzoglou, R. Escourrou, A. Chatzigeorgiou, and I. Stamelos, "A Metric Suite for Evaluating Interactive Scenarios in Video Games: An Empirical Validation", *35th Symposium on Applied Computing (SAC' 20), ACM, Brno, Czech Republic, 30 March —3 April 2020*, 10 pages.

1. Introduction

Games is a special category of software, which is highly pervasive in everyday life of young people and forms a very strong industry. Due to their popularity, and the inherent technical challenges in their development, software engineering for computer games is a rapidly growing research field that attempts to address domain-specific challenges [3]. One of the major differences of games compared to traditional software products is that games' popularity is not related only to the functionality that they provide, but mostly to the satisfaction/enjoyment that they offer to their end-users. Naturally, one of the main key-drivers of video games development is to be entertaining [5]. In the literature, one can identify various research efforts that aim at underlining the main factors that lead to user satisfaction, and consequently entertainment. For example, Ham and Lee [9], and Paschali et al. [20], explored the importance of seven high-level game characteristics – namely: Scenario, Graphics, Sound, Game Speed, Game Control, Character, and Community – in users' satisfaction, through two independent surveys. Based on the results of the most recent study Scenario, Character Solidness and Sound have proven to be the most important factors that influence user satisfaction [20]. Therefore, game development teams should focus on improving these game characteristics, so as to boost games' success, by setting non-functional requirements related to them. Nevertheless, these factors are rather vague and their quality assessment has, until now, not received significant attention. From studying the literature one can detect sets of heuristics or metrics for some of them (see Section 2), but not in-depth quality models, like in traditional software engineering.

To this end, in this study we focus on one of the aforementioned users' satisfaction factors, i.e., the Game Scenario, and provide a model that can be used for assessing the quality of interactive scenarios. The reason to focus on game scenarios is that this characteristic is partially covering the degree to which characters are introduced and interacting. Therefore, by modeling game scenarios we are covering two out of the three most influential game satisfaction factors. We note that for this reason during our modeling, special emphasis is placed on the aspects of scenario design that represent game characters (see Character Model in Section 3.1). The proposed model, first identifies fine-grained scenario characteristics that are related to user satisfaction (by reviewing the literature), and then proposes metrics for quantifying them. To evaluate the validity of the proposed model we conducted a

case study on three interactive scenarios, which have been evaluated by 25 assessors. The main contribution of this study is that, to the best of our knowledge, this is the first study that proposes an objective model for assessing the quality of interactive game scenarios.

The next sections are organized as follows: in Section 2, we present: (a) related work on assessing users' satisfaction; (b) background work on the scenario characteristics that have been associated with user satisfaction, and (c) scenario representation approaches. In Section 3, we present the proposed model, whereas in Section 4 the case study design that has been used for its validation. The results of the validation are presented in Section 5, and discussed in Section 6. Finally, Sections 7 and 8, we present threats to validity and conclude the paper

2. Background Information

2.1 Related Work

Computer games are created in order to entertain their players. According to Ampatzoglou and Stamelos [3], some of the main non-functional requirements of video games are based on specific users' satisfaction factors. According to the same study, the research direction dealing with NFR is the most active one in games' engineering research. The most common way for assessing users' satisfaction is through the use of heuristics, whereas approaches that estimate users' satisfaction from metrics are limited. The rest of the section is organized along these two axes, i.e., heuristics and metrics.

Game Heuristics. Jacobs and Ip organized the preferences (45 heuristics) of rally gamers in nine categories, namely graphics, sounds, technical realism, licensing, structure and challenge, stages and cars, online features, multiplayer features and extras items. Additionally, they suggest that these categories can be used during game design and development, as sources of new requirements [10]. Furthermore, Weibel et al. performed a survey in order to compare the differences from playing online games against human- and computer-controlled opponents. The results of the survey suggest that this factor (i.e. playing against human or computer) could lead to different results in terms of presence, flow and enjoyment [38]. Desurvire et al. formed a set of heuristics for evaluating the playability of games, and classified them into four categories, namely game play, game story, game mechanics and usability [6], [7]. These studies provide specific heuristics, e.g., "Player understands the story line as a single consistent vision", and evaluated their relevance with 54 gamers. In a similar context, i.e., the field of game usability, Pinnelle et al. investigated how the game genre is related to its usability, and the players' ability to learn, control and understand the game interface [24]. In a similar line of though, Piselli et al. concluded that users' in-game enjoyment is related to both gamers' performance and game complexity [25]. Finally, Rookhuiszen and Theune studied the differences between playful and serious instructions, with respect to the entertainment that they offer to the gamer and their efficiency (i.e., clarity and offered guidance). The results suggested that there are differences between the efficiency of serious and playful instructions, whereas no differences have been identified, concerning users' enjoyment [26].

Game Metrics. In contrast to the plethora of heuristics for game evaluation, we have been able to identify only three studies that introduced metrics for a similar purpose. Specifically, Ferreti et al. suggested that the interactivity (number of interactive events) and fairness (guarantee that all players have the same chance of winning, regardless of their subjective network conditions), consistency (a shared view of the game state among all the engaged players) and scalability (assurance that the number of simultaneous players as well as their geographical distribution will be properly scaled) dimensions, should be assessed when designing on-line games [18]. Additionally, Wattimena et al. quantified gaming experience through network metrics. In particular they used measured ping and jitter value to assess the perceived quality of an interactive First-Person Shooter (FPS) game, namely Quake IV. The results suggested that both metrics are highly correlated to quality [37]. Finally, Ampatzoglou et al. proposed the assessment of user satisfaction from 3D scenes, in games and virtual environments, through metrics. Among the metrics, the results of a pilot experiment suggested that the use of advanced texture effects and the number of entities in a scene are the most useful metrics [2].

2.2 Scenario Representation Approaches

Although techniques and rules that lead to an effective way of representing stories have been developed for over a hundred years (e.g., books, movies, etc.), in the domain of video games, they have only quite recently attracted the attention of researchers. In this paper, we focus on scenario representation based on character models and flow charts, as proposed by Paschali et al. [21].

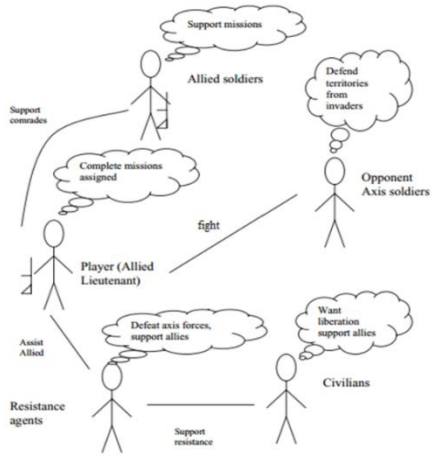


Figure 1: Example character model from the Medal of Honor [25]

A ***Character Model*** is a diagrammatic representation of the characters that are involved in a story/scene, along with their interactions, as described by Rolfe et al. [26]. For example, in [26] the authors describe a scene from the Medal of Honor game, with the following character model (see Fig. 1). The main notations of the diagram are the characters of the game (stickmen – e.g., Allied Soldier), their interactions (continuous lines – e.g., the Player is fighting with Opponent Axis Soldiers), and their high-level goals (though bubbles – e.g., the goal of the Civilians is to be liberated and receive support by allies).

Flow Charts may often be included as part of the game design document, similarly to those of traditional software engineering. Flow charts are diagrams that represent an algorithm, workflow or process, showing the steps as boxes, and their sequence of execution by connecting them with arrows. In game development, flowcharts are used to track [30]: (a) players' navigation of out-of-game menu options (e.g., starts a new game or loads a saved one), and (b) areas the players progress to and from in the game, particularly in level-based games.

Beyond these most obvious applications, flowcharts can be quite useful for visually representing the results of any decision players may take during a game [30]. In some game genres (e.g., MMOG - Massively Multiplayer On-Line Games) interactivity is a distinguishing feature and an attraction for gamers, since participants can change the state of affairs with their actions. In such games, due to the dynamic flow of events, gameplay can be resembled to the execution of an algorithm, where elementary actions are defined by game rules, rendering the flowchart a fitting means for their representation [30][34].

2.3 Scenario Characteristics

In this section, we present the different aspects of scenarios that have been evaluated so far in the literature. The goal of this section is the identification of characteristics, which can be used in the evaluation of video games and more precisely in the evaluation of their scenario. An overview of characteristics is presented in Table 1. For each characteristic, we denote its frequency and provide pointers to the studies that have employed them for scenario evaluation purposes. The results of Table 1 provide only a coarse-grain estimation of the importance of each characteristic, since they have not been obtained through a systematic process.

Table 1: Overview of Scenario Characteristic

Characteristic	Freq.	Reference
Narrative Level	4	[14], [15], [12], and [33]
Re-playability	1	[12]
Interactivity	7	[14], [15], [12], [33], [35], [32], and [29]
Characters' Interaction	5	[14],[15],[13], [12], [35], and [29]
Content	6	[14],[15],[13], [12], [22], and [33]
Coherence	1	[22]
Originality	1	[22]
Achieved curiosity	3	[28], [29], and [35]
Immersion	6	[36], [12], [33], [28], [29], and [35]
Desirability	2	[12] and [32]

Narrative Level has been defined as a scenario characteristic that aims at evaluating the extent to which a game contains an appropriate introduction, sub-goals and ending [12]. Most of the heuristics used for assessing this characteristic are based on elements such as the game interface, mechanics, and gameplay. Specifically, Sweetser and Wyeth [33] have compiled a concise model of enjoyment in games, structured by flow, based on the aforementioned heuristics. According to Macvean and Riedl [14], [15] there are five rules or heuristics (as originally defined by Koeffel et al. [12]) that should be followed in order to construct an appropriate narrative structure. For example, they suggest that a game should be: (a) clear in the way it defines failure conditions, (b) consistent and (c) respond to the user's actions in a predictable manner.

Re-playability has been defined by Koeffel as the ability of a game to create the desire to the user to play it more than once [12]. In the same study, Koeffel dedicated more than one heuristic for highlighting the importance of re-playability (e.g., the first gaming experience should not disappoint the user, but encourage him/her to accomplish its goals). This property is frequently referred in the gamers' community as "easy to play, hard to master", based on the Bushnells' law [4]. We note that although re-playability is probably influenced by the overall experience offered by the game to its users, in this paper we explore the contribution of scenario in the overall desire of the player to play the game again.

Interactivity is also defined by Koeffel [12] as the ability of the game to make players feel that they have the control of the characters and an impact onto the game world. Specifically, players' actions should matter and they should shape the game world [14],[15]. By surveying the literature, we have been able to identify two almost synonym quality attributes for interactivity: Control [33] (i.e., players should feel the sense of control over the actions and feelings of their characters) and Effectance [35], [29] (players should be able to feel that they are affecting the game world). Finally, Schoenau [32] highlights the importance of interactivity, by suggesting that it is the most important requirement for continuing playing the game.

According to Macvean [14], [15], Khan [11] and Koeffel [12], game ***characters*** must be interesting and relevant to the story. Despite the fact that according to Paschali et al. [20] and Ham et al. [9], game characters and game scenarios constitute different game satisfaction factors (see Section 1), in this study, we treat them uniformly since the characters are main units of the game plot. In some cases, the whole game scenario is built around characters, and their decoupling is very difficult and could pose a threat to validity. Lankoski [13] highlighted the importance of believability that supports that: the game is believable when players are able to interpret the game events and character actions without much effort. In a similar line of thought, virtual characters must not damage user's illusion by irrational behavior or poor response to user input [35], [29].

Story Content is an ingredient, which could drive to interesting stories [36]. In particular, according to Sweetser, users are satisfied if each part of the game scenario story is smoothly located in the context of the overall story [33]. Additionally, Macvean et al. suggest that there are two criteria for assessing game flow, namely: (a) the game contains interesting and varied sub-goals, and (b) the game is of appropriate difficulty [14], [15]. Furthermore, the same studies underline that stories should be modular in nature and that their content at each location should be fit to the overall narrative. The same path is followed by Peinado [22] and Lankoski [13], who characterized game quality in eleven components including features which are relevant to content and flow, such as moral choices and optional side quests [13]. Finally, Koeffel suggests that players shouldn't be burdened with tasks that are not deemed as important [12.]. Similarly to traditional software engineering, ***Coherence*** describes how well a sequence of events is linked. This aspect of story quality is discussed by Peinado who discusses the linking between concepts, data-type and object properties [22]. The characteristic of Linguistics, as introduced by Peinado describe how accurately the in-game texts are written. For example, every event is related to a rationale cause and effect [22]. Finally, in the same study Peinado discuss story's ***Originality***, which describes how different a story is from others [22].

Curiosity is achieved when players become absorbed in what will happen next. According to Roth triggering users' curiosity is of

paramount important in the field of entertainment media [29]. Furthermore, Roth [30] and Vermeulen [35] suggest that curiosity can be measured through user responses to interactive stories. Suspense is achieved when the players develop hopes and expectations. This characteristic also reflects to uncertainty about the progress [29], [30], [35]. ***Immersion*** is achieved when players feel deeply, emotionally involved in the story without the sense of time [Sweetser and Wyeth]. The story emotionally transports the player into a level of personal involvement: scare, threat, thrill, reward, and punishment [12], or it could be described as the difference between the utility that participant feel that gains when wins or loses with the dimension of intensity [36]. This type of dedication is also emphasized by Roth et al. and Vermeulen [29], [30], and [35]. Players should feel the ***Continuous Desire*** to not stop playing the game [12], which should be driven by the game itself [32]. We note that the difference between re-playability and continuous desire is very thin, and therefore misinterpretations are possible. Nevertheless, we tried to separate the two terms by explaining that re-playability is the will of the player to play the game again (after closing it), whereas continuous desire is the will of the player to not stop playing the game (i.e., not close it).

4. Proposed Game Scenario Quality Model

In this section, we describe the approach that we propose for predicting users' satisfaction from game scenarios. As a basis for model application, we consider the tailored version of flow charts and character models, as proposed by Paschali et al. [21]. To tailor flow charts for designing game scenarios, we propose the use of some additional notations that are useful for the desired representation. Under this tailored representation, the story will be divided into three parts (Exposition, Rising Action, Climax, and Endings), as dictated by the narrative structure. The notations used for the tailored flow charts are:

- ***Rectangles/Actions*** represent sequences of actions or events during which the player is passive. These sequences are used to set up the next situation or show the consequences of successful (or unsuccessful) completion of previous tasks.
- ***Choice/Fork*** represent a "free play area" in-side the story, i.e., choice. The players can make choices which will impact the unfolding of the story or other players. As a choice we classify any action of the player that can alter the flow of events; e.g., solving a puzzle, can lead to unlocking a completely new path in the game flow, which would not be revealed to the player, if he/she would not be able to solve the puzzle or if he/she had lost the battle.
- ***Filled rectangles/Goals*** are used to show the goals in the story.
- ***Ovals/Ends*** denote the endings and starts of the story. The possible different endings of the story are denoted with white for "happy ending", and black for "bad ending", whereas the start of the story is denoted with a grey oval. In the special case of games with only one type of ending (e.g., the game finishes and the player is provided with a score, so as to compare it with other players), this end is denoted as a "happy end". For games that do not have an obvious ending, e.g., SIMS, there is no ending.
- ***Arrows*** are used to show the direction of the flow in the story.
- ***Swimlanes*** denote the different parts of the story (Exposition, Rising Action, and Climax).

The model was supported by a tool, namely UMBRA, which produces the outcome presented in Fig. 2.

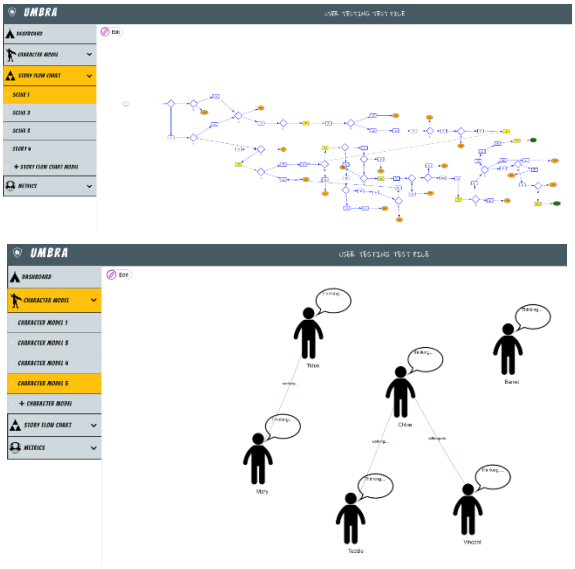


Figure 2: Designing Game Scenarios in UMBRA

The proposed model is hierarchical and has two levels (see Fig. 3): the first represents the scenario characteristics described in Section 3.2. The last level corresponds to the metrics that we propose for quantifying them. At this point it is necessary to clarify that the proposed metrics can be directly calculated from flow charts and character models, which according to Section 3.1 are among the most established techniques for representing game scenarios. Although, at this point we do not provide a tool for the calculation of metrics, the automation of the process is trivial and is included in the authors' future research plan. In Fig. 3, quality characteristics are divided into two categories, with respect to whether they can be quantified by metrics obtained from flow charts or character models.

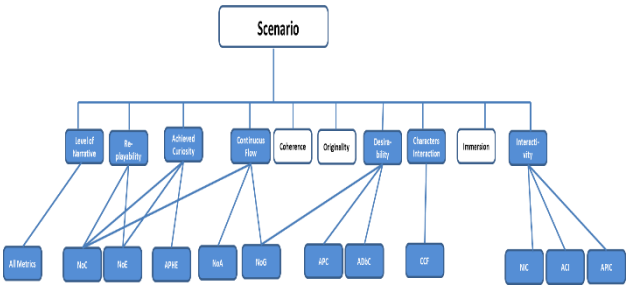


Figure 3: Scenario Evaluation Model

In particular, quantifiable quality characteristics are denoted with a transparent rectangle, whereas quality characteristics that cannot be quantified are denoted with a filled rectangle. In particular, we believe that Story Coherence cannot be measured from flow charts, in the sense that it represents the conceptual relevance of game actions and not their structural interconnection. Therefore, in order for a metric to be able to capture coherence, a textual representation of game scenarios should have been employed (e.g., through UC descriptions). Similarly, Story Originality, which represents the extent to which a story is different from other stories, cannot be evaluated in isolation. Therefore, in order for such a metric to be created, data for other stories are needed, whereas at the same point, again, conceptual information is needed. Finally, in our point of view, players' Immersion into a game is a more high-level characteristic, which is achieved by ensuring

other properties, for example, in order for a player to be immersed into the story he must be curious about the outcome of his/her actions, he/she must be interested in the characters, the narrative structure must ensure the interest of the player in all parts of the game, etc. To this end, we believe that a single metric is not able to accurately predict this aspect of quality. Next, we will present the metrics, which can be calculated from flow charts and character models. The rest of the section is organized based on scenario characteristics. Although, level of narrative is the first quality characteristic in Fig. 3, we discuss it in the end of the list, since it makes use of all other metrics. As it can be observed from the following paragraphs, each quality characteristic is associated with multiple metrics.

Re-playability: Based on the definition of the re-playable characteristic, we assume that a gamer would be willing to play a game for a second time, if the game provides multiple choices and multiple endings. In particular, we assume that games whose outcome is not heavily based on players' input will not make sense to be played for a second time. On the other hand, a game which offers various alternatives can be considered re-playable, in the sense that the gamer might be interested in exploring all possible endings of the game. To this end, we base the evaluation of the re-playability quality characteristic on two metrics, defined as follows:

```
Number of Choices, NoC = Cardinality (Choices)
Number of Ends, NoE = Cardinality (Ends)
```

Interactivity is a quality characteristic related to the response of a game as a reaction to the movements/actions of the players, i.e., the sense of control that the player has inside the game. Obviously, this is an attractive feature for the gamer and can lead to high engagement. In contrast, the absence of this attribute could make the gamers feel demotivated for playing the game. By taking the above into consideration we evaluate the interactivity characteristic with the usage of three metrics, related to choices. The metrics are defined as follows:

```
Number of Important Choices, NIC = Cardinality(ChoicesWeight =5)
AVG Choice Importance, ACI = AVG(ChoicesWeight)
AVG Paths after Important Choices, APIC = AVG(Out_Degree(ChoicesWeight =5))
```

Characters' Interaction: Isolated characters cannot play an important role in the plot of a game. On the other hand, characters that are connected and heavily interact with each other are expected to make the game more interesting and facilitate user satisfaction. The proposed metric) for quantifying characters' interaction is based on the object-oriented coupling factor (CF) metric, calculated as follows:

```
Characters CF, CCF = Cardinality (Edges) / (Nodes 2)
```

Content: The content is an important aspect of the game's story. In order for a story to provide a rich content to the gamer, it must provide him/her many possible actions, choices and goals. We evaluate the content with three metrics, which are defined as follows:

```
Number of Actions, NoA = Cardinality (Actions)
Number of Choices, NoC = Cardinality (Choices)
Number of Goals, NoG = Cardinality (Goals)
```

Achieved Curiosity: To ensure that the curiosity of the gamer, and provided suspense, is safeguarded during the game, we linked this quality characteristic with metrics related to the possible endings of the game and the number of choices. In addition to that, we believe that the number of paths from which a happy ending can be reached are increasing the suspense of a game, in the sense that a wrong action does not necessarily lead to a bad ending. The fact that two of the used metrics for assessing the achieved curiosity overlap with those of re-playability is due to the relationship between the two quality characteristics, i.e. a game that raises the curiosity of the user is more probable to be played again. Specifically, we linked the degree to which curiosity is achieved are defined as follows:

```
Number of Choices, NoC = Cardinality (Choices)
Number of Ends, NoE = Cardinality (Ends)
AVG Paths to Happy Endings, APHE = AVG(In_Degree(Endshappy))
```

Desirability is dealing with the reasons that can make the player to desire to continue playing the game. In order for a gamer to play a game for long periods it must offer many actions, and have a large duration. In addition to that, we believe that also the possible different paths that the game can take are also related with the will of the gamer to continue playing the game, as well as the frequency of choices (important or not). Therefore, in order to assess the desirability quality characteristics, we use three metrics, which are defined as follows:

```
Number of Actions, NoA = Cardinality (Actions)
AVG Paths after Choices, APC = AVG(Out_Degree(Choices))
AVG Distance between Choices, ADbC = AVG(Distance (Choices[i], Choices[i+1])), ∀ Choices
```

Level of Narrative: In order to investigate if the game scenario follows the desired narrative structure as described by the Freytag's pyramid (see Section 3.1), we need to take into account all the aforementioned metric scores in the five phases of the scenario (i.e., Exposition, Rising Action, Climax, Falling Action, and Conclusion). In particular, all metrics are expecting to increase between: (a) the Exposition and the Rising Action phase, (b) the Rising Action and the Climax phase, and decrease between the Climax and the Falling Action phase. Therefore, for each metric, we count how many transitions conform to the aforementioned rules. The idea of calculating a metric, based on a set of other metrics has been inspired by the reliability property, as introduced in the 1061 IEEE Standard for Software Quality [1].

5. Empirical Validation

In this section we present the design of the case study [31] that we have performed for investigating the accuracy of the proposed mapping between metrics and quality characteristics. In particular, we used three interactive scenarios, in the form of interactive books (or game books), so as to ensure that the rest game satisfaction factors (e.g., graphics, sound, etc.) do not confound the evaluation. The term interactive book is used for books that allow the reader to participate in the story by making effective choices. The narrative branches along various paths through the use of numbered paragraphs or pages. The main reason that we performed a case study rather than another type of empirical evaluation (e.g., survey [23] or experiment [39]) is that we wanted to test our model in practice, using real scenarios and evaluators. In addition, although we have filtered out as many confounding factors as possible, we cannot argue that this study offers the level of control required by experiments.

Objectives and Research Questions. The goal of this case study is to evaluate the validity of the proposed model. To achieve this goal, we decompose it to two research questions:

[RQ_1] What is the rate of agreement of interactive scenarios' evaluators, with respect to the seven scenario characteristics (level of narrative, re-playability, interactivity, characters' interaction, Content, achieved curiosity, and desirability) involved in the proposed model?

Answering RQ_1 will allow us to understand which scenario characteristics are uniformly assessed by independent evaluators. A positive answer to this question will mean that the specific quality characteristic is perceived in a similar way by the majority of evaluators, and therefore an accurate prediction is possible. On the other hand, if evaluators do not agree on their evaluation on a specific characteristic, this will automatically mean that any possible metric cannot achieve a decent accuracy while trying to predict this. Therefore, the answer to this question can validate if the proposed model has correctly identified and defined scenario quality characteristics.

[RQ_2] What is the accuracy of the proposed metrics in predicting the satisfaction that a user gets from an interactive scenario, with respect to the seven scenario characteristics (level of narrative, re-playability, interactivity, characters' interaction, content, achieved curiosity, and desirability) of the proposed model?

Answering RQ_2 will lead to either confirming or rejecting the mapping between metrics and quality characteristics. A positive answer for a pair of metrics and quality characteristics will mean that using these metrics it is possible, at design-time, to predict the user satisfaction that a wide audience will get from a specific scenario. On the other hand, a negative answer will mean that this quality characteristic needs further investigation so as to extract other metrics that can predict its value at design-time.

Case Selection and Unit Analysis. This study is a holistic multi-case study, in which each interactive scenario is both a case and a unit analysis. In particular, as cases we selected three interactive books of small size, namely: "Journey Under the Sea", "Underground Kingdom", and "The Cavern of Doom". The books are part of the same interactive book family (i.e., Choose your own adventure) and therefore we expect that they do not have major differences in the writing quality, style, and theme. Each book has been read by twenty (20) subjects in a timeframe of 15-20 minutes (i.e., one hour for the whole case study). Next, each participant was given a data collection form in which he/she filled in his/her opinion about the books, as described in Section 5.3. The subjects have been randomly selected and were of different ages, nationalities and educational level.

Data Collection & Analysis. For each investigated scenario quality characteristic (i.e., level of narrative, re-playability, interactivity, characters' interaction, Content, achieved curiosity, and desirability) the subjects have been asked to rank the three stories from the one that was the most satisfying to the least satisfying. We clarify that subjects were introduced to the investigated concepts prior to filling the forms through a presentation given by the first author. The final dataset was consisted of twenty (20) lines—one for each subject, and fifteen (15) columns/variables, characterized through an id [V1] as shown below:

Ranking of stories with respect to:

[V2] the level of narrative
[V3] their re-playability
[V4] their interactivity
[V5] the level of interaction between their characters
[V6] the level of content
[V7] the level of achieved curiosity
[V8] their desirability

Predicted ranking of stories with respect to:

[V9] their level of narrative
[V10] their re-playability
[V11] their interactivity
[V12] the level of interaction between their characters
[V13] the level of Content
[V14] the level of achieved curiosity
[V15] their desirability

Variables [V2] – [V8] represent the ranking of stories as provided by the evaluator, whereas variables [V9] – [V15] represent the predicted ranking, based on metrics. To obtain the predicted ranking the three scenarios must be compared in pairs, with respect to the given quality characteristic. To compare two scenarios with respect to one quality characteristic the following steps have to be performed:

- Calculate all metrics associated to the quality characteristic under study, for both scenarios.
- Identify which scenario excels with respect to each metric.
- Select as optimal the scenario that excels for the majority of metrics. In case of tie we consider the two scenarios as equivalent with respect to the specific quality characteristic.

We selected to use the aforementioned process for combining metrics, rather than a weighted sum or another mathematical aggregation function, due to the diversity of the examined metrics. To answer the research questions stated in Section 5.1, we followed the process below:

Agreement between subjects. To check the agreement among the ratings obtained from all subjects, we performed the inter-rater reliability analysis. We calculated the average ICC, which represents the average correlation among all raters, for variables [V1] – [V8].

Metrics prediction accuracy. In order to quantify the accuracy of the proposed metrics in predicting the satisfaction obtained from interactive scenarios, we used correlation analysis [Field]. The decision to apply a correlation analysis (Kendall's tau rank correlation) is based on the 1061 IEEE Standard for Software Quality Metrics Methodology [1], which suggests that a sufficiently strong correlation "determines whether a metric can accurately rank, by quality, a set of products or processes (in the case of this study: a set of methods)". Kendall's Tau Distance is used to quantify the similarity between responses by taking into account the distance between two orderings. The distance between orderings is calculated as the pair-wise differences between the two lists. When comparing two orderings of length 3, the minimum distance is zero (0)—both orderings are exactly the same, whereas the maximum is three (3)—one ordering is the reverse of the other. For example, the distance between {A B C} and {C B A} is 3, because the pairs {A B}, {A C}, {B C} are inverted in the second ordering. Kendall's tau rank correlation coefficient (γ) is used to measure the association between multiple rankings:

$$\gamma = \frac{numberofconcordantpairs - numberofdiscorantpairs}{0.5 * n * (n-1) * N}$$

For our study, n = 3 (number of items in the list) and N = 20 (number of raters). The process is applied to pairs of variables that correspond to the same scenario characteristic (e.g., [V2] / [V9], [V3] / [V10]).

6. RESULTS

To facilitate the readability of the case study results, in this introductory section, we present the metrics calculated from for every story (see Table 2). The flow charts and character models for all evaluated stories are presented online. The rest of the section is organized by research question. We note that due to the fact that this is the first study that proposes metrics for the quantification of game scenario quality characteristics, we do not have an extensive comparison to related work. For the rest of this section we prefer to use the term 'game scenario' rather than the term 'book scenario' for referring to the units of analysis of our case study, so that discussions match the target context, i.e., game development.

Table 2: Story Evaluation

Story	NoC	NoE	NIC	ACI	APIC	CCF
1	46,0	42,0	0,19	3,28	2,06	0,40
2	22,0	21,0	0,05	2,95	2,11	0,52
3	18,0	17,0	0,06	2,77	2,15	0,43
Story	**NoG**	**APHE**	**NOA**	**APC**	**ADbC**	
1	5,0	6,0	9,0	2,0	0,2	
2	9,0	2,0	47,0	2,0	1,3	
3	9,0	1,0	60,0	2,2	2,0	

6.1 Inter-rater Agreement

Regarding RQ1, we calculated the average ICC correlation among the ratings of the 20 evaluators, by scenario quality characteristic. The results are summarized in Table 3. In order to interpret the values obtained by the correlation analysis, we use the threshold provided by Marg et al. [16] (e.g., correlation coefficients between 0.7 and 0.9, characterize very strong correlations). To visualize the strength of correlation, in Table 3, very strong correlations are denoted with dark grey cell shading, whereas strong correlations with light grey shading.

Table 3: Inter-rater Agreement

Characteristic	ICC
Level of narrative	0.47
Re-playability	0.92
Interactivity	0.78
Characters' interaction	0.72
Content	0.64
Achieved curiosity	0.87
Desirability	-1.81

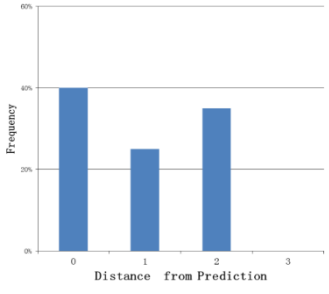
Based on the results of Table 3, we can claim that the re-playability, interactivity, characters' interaction, and achieved curiosity characteristics are uniformly evaluated in game scenarios by all evaluators, in the sense that individual ratings are strongly correlated. Additionally, concerning Content and level of narrative the ICC is strong and therefore, there is enough agreement between raters, so as to assume that a model is able to adequately predict user preferences. On the other hand, desirability is having a strong negative correlation among raters. Therefore, we believe that no model is able to reach adequate prediction accuracy. Thus, any results concerning desirability should be treated with caution.

The levels of inter-rater agreement are expected to be influenced by two main parameters: (a) the abstractness of the concept, and (b) the accuracy with which we explained the participants each concept. Although the first parameter can be characterized as subjective, we believe that some quality characteristics are by nature concrete and easy to perceive—e.g., Would you like to play the game again (re-playability)?—whereas others more obscure—e.g., discrimination of game scenario phases (Exposition, Rising Action, Climax, Falling Action, and Conclusion) needed to assess the narrative structure. Concerning the second parameter, an objective measure of how established a concept is expected to be, is the number of primary studies in which the concept is explained. For instance, a concept that is highly studied (e.g., Interactivity) has provided us with more examples of its meaning, and thus it was easier for us to explain it to the subjects. Therefore, we would expect the frequency of quality characteristics in Table 1 (see Section 3) and the Inter-rater agreement (see Table 3) to be related. However, some quality characteristics like re-playability or achieved curiosity, although discussed only in 1 study, they exhibit a high inter-rater correlation. A possible explanation of that is their ease of perception (see first parameter). The rest quality characteristics are more or less ranked in a similar order. Thus, this second parameter can be a solid explanation of our failure to explain the term desirability to subjects, i.e., it is discussed only in one study. The most common reason of misinterpretation was its conceptual relevance to re-playability, which describes the will to play/read the game/book again after you close it ones, whereas desirability describes the will to not close the game/book.

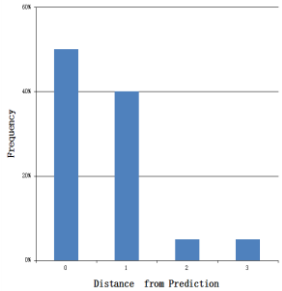
All scenario characteristics, except desirability, present an adequate level of agreement between raters, and consequently the accuracy of the proposed model can be evaluated on them. Desirability needs to be redefined as a quality characteristic, because it is not uniformly perceived from independent raters.

6.2 Metrics Prediction Accuracy

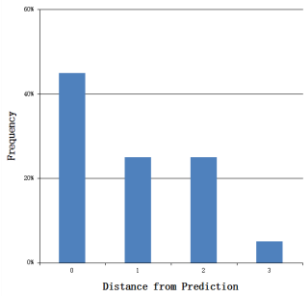
Concerning RQ2, we first present some descriptive statistics on the distance between the rating of evaluators and the metrics outcome. In particular, in Fig. 3.a – 3.g, we present the bar charts for every scenario characteristic: in the x-axis we present all possible values of the Kendall's Tau Distance (ranging from 0—absolute agreement to 3—absolute disagreement), whereas in the y-axis the percentage of their frequency over the total number of evaluators.



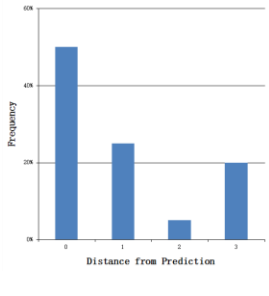
(a)- Level of narrative



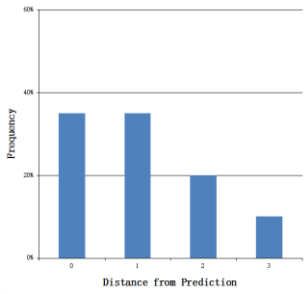
(b)- Re-playability



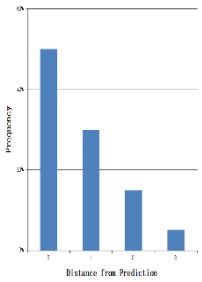
(c)- Interactivity



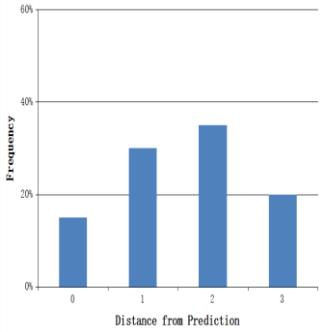
(d)- Characters' interaction



(e) –Content



(f) – Achieved Curiosity



(g) - Desirability

Figure 3: Descriptive Statistics on Prediction Accuracy per Scenario Characteristic

The results of Table 4 provide a more fine-grained assessment. Specifically, in Table 4, for each scenario characteristic, we present the concordant pairs, the discordant pairs, and the Kendall's tau rank correlation coefficient for all books. In order to ease the parsing of Table 3, we highlight with dark grey color the strong correlations, whereas with light grey the moderate correlations. We note that since the prediction that is attempted through this study is considered as highly ambitious (i.e., assess a purely subjective variable—user satisfaction—from objective ones), we consider correlations > 0.4 as adequate. From the results of Table 4 we can observe that the proposed metrics are having a strong predictive power (i.e., correlation higher than 0.4) on scenarios' re-playability, achieved curiosity, and interactivity. These quality characteristics are among the four quality characteristics for which the assessors had a very strong rate of agreement. This is a rather intuitive finding in the sense that concerning characteristics for which the opinion of evaluators is not uniform, any model is highly unlikely to present a strong predictive power. The most evident example of this category is scenarios' desirability for which evaluators had very diverse opinions (correlation coefficient: -1.81), and the predictive power of our metrics was almost zero (correlation coefficient: -0.06). Another interesting finding by opposing the results of Table 3 and Table 4 is that the proposed metrics for narrative structure are the 4th more strongly correlated ones to experts' opinion, regardless the fact that with respect to the correlation among evaluators, it was ranked 6th. This fact implies that the predictive power of the proposed metrics is lowered due to the inherent diversity of the aspect that we aim at predicting.

Table 4: Metrics Predictive Power

Characteristic	Concordant pairs	Discordant pairs	Kendall's Tau
Level of Narrative	41	19	0.37
Re-playability	47	13	0.56
Interactivity	42	18	0.40
Characters' Interaction	41	19	0.36
Content	39	21	0.30
Achieved Curiosity	45	15	0.50
Desirability	28	32	-0.06

In order to present the extent to which the predictive power of the metrics is close to the highest possible predictive power, in Table 5 we present: (a) the name of the quality characteristics, (b) the predicted ordering of stories, and (c) the most frequent ordering of stories based on evaluators' opinion. For (b) and (c) we present the percentage of evaluators that selected the corresponding ordering of stories and the Kendall's tau rank correlation coefficient for this ordering. The results of Table 5 suggest that for all quality characteristics, apart from desirability, our metrics pointed out to the most popular story ordering. Nevertheless, in none of the cases more than 50% of the evaluators had a common perception of the ranking.

Table 5: Metrics Predictive Power

	Predicted Ordering of Stories			Most Frequent Ordering of Stories		
Characteristic	Stories Order	Pct.	K. Tau	Stories Order	Pct.	K. Tau
Level of Narrative	C-B-A	40%	0.37	C-B-A	40%	0.37
Re-playability	A-B-C	50%	0.56	A-B-C	50%	0.56
Interactivity	A-C-B	45%	0.40	A-C-B	45%	0.40
Characters' Interaction	B-C-A	50%	0.36	B-C-A	50%	0.36
Content	B-C-A	35%	0.30	B-C-A	35%	0.30
Achieved Curiosity	A-B-C	50%	0.50	A-B-C	50%	0.50
Desirability	C-B-A	15%	-0.06	A-B-C B-A-C B-C-A	20%	0.07

The metrics proposed for Re-playability, Interactivity, and Achieved Curiosity are strongly correlated to the corresponding quality

characteristics. Metrics for Level of Narrative, Characters' Interaction and Content are still in need of improvement (moderate correlation). As expected, Desirability is not possible to be predicted by any metric, since the raters had a different understanding of the concept.

7. Discussion

The results described in this paper can be considered important for both game researchers and practitioners. Concerning practitioners, we expect that the proposed model (see Section 4) will help them to improve the game design process by:

- ***Evaluating the quality of the game scenario in early development stages***. Similarly, to software quality assessment, early quality indicators are cost-efficient for software development companies. Therefore, the use of metrics that assess re-playability and achieved curiosity (the quality characteristic that are more accurately assessed by the proposed model) can lead the game design team to changes in the scenario (e.g., more choices and endings) that can potentially increase game popularity, with minimized cost. In particular, the fact that the proposed evaluation is performed on the design phase (i.e., before the start of the implementation) can help in avoiding changes after the usability testing phase, which would be costlier for the company.
- ***Providing useful design modeling notations***. According to the literature, the game development processes are usually agile (i.e., in many cases no analysis or design artifacts are being produced), and in many cases, no specific development processes are followed. The results of this study indicated that some design artifacts (i.e., flow charts and character models) can be useful, not only for designing purposes, but also for quality assessment ones. Therefore, based on this, we highly encourage game designers to develop such artifacts, since the obtained benefits are two-fold.

On the one hand, based on the results of this study we have been able to extract important implications for researchers and identify interesting future work directions, as follows:

- ***Assessing external quality attributes from internal ones***. Despite the fact that the prediction of external quality attributes (i.e., user satisfaction) from internal ones is in general an ambitious goal, the results of our study suggest that is a feasible target for moderate to strong correlations. Therefore, we consider the further exploration of this field as an interesting direction.
- ***Improvement of existing model.*** Although the majority of the examined quality characteristics have been sufficiently predicted by metrics, in some of the cases the proposed model needs refinements. In particular, the definition of the desirability characteristic should be revisited, since it was not uniformly perceived by the evaluators of our study (very low intra-class correlation, see Table 4). Also, since the achieved predictive power for some quality attributes is only moderate (e.g., Level of Narrative, Characters' Interaction and Content—see Table 5) further investigation in the selection and aggregation of the associated metrics is needed.
- ***Replication of the study***. As in any empirical study, so as to increase the validity of the results, a replication of the process is required, especially with respect to the application on real games. Therefore, we encourage the repetition of the case study with a higher number of evaluators and scenarios.
- ***Development of similar models for other quality characteristics***. Scenario is not the only factor that is related to user satisfaction. Therefore, we highly recommend researchers to build and validate similar models for other satisfaction factors (e.g., graphics, controls, etc.)

8. Threats to Validity

In this section, we present construct, reliability, external, and internal validity threats for this study. Construct validity reflects the extent to which a phenomenon under study is represented by the research setup. Reliability is associated to the ability of others researchers to repeat the same process, collect data and reach the same results. External validity deals with issues rose while generalizing the findings of the study. Finally, internal validity is related to the identification of confounding factors, i.e., factors other than the independent variables that might influence the value of the dependent variable.

Internal Validity. The proposed study attempted an association between high level quality characteristics (such as re-playability and interactivity) and particular metrics extracted from game flow chart representations (such as number of choices or endings). Obviously, one cannot claim that the characteristics of interest have a one-to-one mapping with the selected metrics, a fact which might raise threats to the internal validity of the study, as other, possibly omitted metrics might affect the investigated relations.

Construct Validity. Despite the fact that the goal of the study is to develop a method for assessing game scenarios, the objects used in this case study have been interactive books and not computer games. This choice poses a threat to validity in the sense that the intended context is slightly different than the one used in the research setup. However, we believe that the choice of interactive books brought an important benefit to this case study, since it eliminated other confounding factors, such as graphics, controls, etc., that could bias the subjects. In addition to that, we believe that a scenario of a game and a scenario of interactive book are very close in their nature, and therefore any possible bias from this is limited. An additional threat to construct validity was that although the stories belong to the same series, they have been written by different authors (implying changes in writing style) and have a different plot. Therefore, some authors might rank the books, not based on their technical quality, but on their personal interest in the topic. Nevertheless, we believe that books of the same series cannot be considerably different, and therefore this cannot hugely influence the results. Finally, to decrease possible bias in favor of the last book that evaluators have read, we shuffled the order with which they were provided the books. Additionally, the selection of smaller books, would not guarantee its similarity to a complex game, and therefore would raise additional threats to validity.

Reliability. The process that has been followed in this case study has been documented in detail in the case study protocol, presented in Section 5. Therefore, the execution of the case study is reproducible by any interested researcher. However, a possible threat to reliability is related to the observed lack of agreement among the participants. This means, that different evaluators might lead to different scenario rankings, and therefore different results. Nevertheless, despite the actual predictive power scoring, the proposed metrics are always achieving top predictability (see Table 5).

External. Concerning external validity, the generalizability of our results from the sample to the population is threatened by the small

sample size of this case study, in terms of evaluators, and books. Therefore, the replication of this study with a larger number of evaluators, scenarios, and in a more realistic context (see implications to researchers—Section 7) would be valuable.

9. Conclusions

Game Scenario has been reported as one of the most important user satisfaction factors in computer games. However, current literature lacks approach for quantifying the satisfaction that users get from playing games. In this path we developed a model that assesses quality characteristics of interactive game scenarios through metrics calculated from design documentation. To validate the proposed model, we conducted a case study on three interactive scenarios, which have been evaluated by 25 participants. From the aforementioned process we have validated that the proposed metrics are able to accurately assess the level of quality for three scenario characteristics (i.e., Replayability, Interactivity, and Achieved Curiosity), and adequately predict it for three more (i.e., Level of Narrative, Characters' Interaction and Content). However, our model was not able to predict the Desirability characteristic, and therefore either its definition or metrics associated with it should be reconsidered. Nevertheless, we need to note that the task of predicting user satisfaction from design artifacts is a hard task, in the sense that in some cases (3 out of 7), even the assessments of evaluators were not in accordance. Based on the results of this study we have been able to provide some useful implications for researchers and practitioners. For example, the proposed model can be used by practitioners in achieving early assessments of their game scenarios (through artifacts produced in the design documents), leading to cost-efficient revisions of games that are not expected to achieve user satisfaction. On the other hand, researchers are provided with various potential future research directions.

References

[1] 1061-1998: IEEE Standard for a Software Quality Metrics Methodology, IEEE Standards, IEEE Computer Society, 31 December 1998.

[2] A. Ampatzoglou, A. Chatzigeorgiou and I. Stamelos, "Graphical Representation as a factor of 3D software user satisfaction: A metric-based approach", 12th Panhellenic Conference, IEEE Computer Society, Samos, Greece, 28 - 30 August 2008.

[3] A. Ampatzoglou and I. Stamelos, "Software Engineering Research for Computer Games: A systematic Review", Information and Software Technology, Elsevier, 52 (9), pp. 888-901, September 2010.

[4] N. Busnhnell, "Bushnell's Theorem: Easy to Learn, Difficult to Master", Wolfshead Online. Retrieved 26 February 2014.

[5] D. Callele, E. Neufeld, K. Schneider, Emotional requirements in video games, in: Proceedings of the International Conference on Requirements Engineering (RE'06), IEEE Computer Society, Minneapolis, MN, USA, 11–15 September 2006, pp. 292–295

[6] H. Desurvire, M Caplan, J.A. Toth, "Using Heuristics to Evaluate the Playability of Games", Proceedings of the Special Interest Group in Computer Human Interaction (SIGCHI'04), Association for Computing Machinery, pp. 1509 – 1512, Vienna, Austria, 24-29 April 2004.

[7] H. Desurvire and C. Wiberg, "Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games - The Next Iteration", Proceedings of the 3d International Conference on Online Communities and Social Computing, Springer-Verlag, pp. 557-566, San Diego, California, USA, 19 – 24 July 2009.

[8] A. Field, "Discovering Statistics using IBM SPSS Statistics", SAGE Publications, 2013.

[9] H. Ham and Y. Lee, "An Empirical Study for Quantitative Evaluation of Game Satisfaction", 2006 International Conference on Hybrid Information Technology, ACM, pp. 724-729, November 2006.

[10] G. Jacobs and B. Ip, "Establishing user requirements: incorporating gamer preferences into interactive games design", Design Studies, Elsevier, 26 (3), pp. 243-255, May 2005

[11] U.A Khan and Y. Okada, "Evolving story and character generation for role-playing games" Workshop at SIGGRAPH Asia (WASA '12), ACM, New York, NY, USA, 59-64, 2012

[12] C. Koeffel, W. Hochleitner, J.Leitner, M. Haller, A. Geven, M. Tscheligi, "Using Heuristics to Evaluate the Overall User Experience of Video Games and Advanced Interaction Games", In Evaluating User Experience in Games, pp 233-256, 2010

[13] P. Lankoski, "Models for Story Consistency and Interestingness in Single-Player RPGs", International Conference on Making Sense of Converging Media (MindTrek '13), ACM, New York, NY, USA, Pages 246, 8 pages,2013.

[14] A. McVean and M. Riedl, "An enjoyment metric for the evaluation of alternate reality games", 6th International Conference on Foundations of Digital Games (FDG '11). ACM, New York, NY, USA, pp 277-279, 2011.

[15] A. McVean and M. Riedl, "Evaluating enjoyment within alternate reality games", ACM SIGGRAPH 2011 Game Papers (SIGGRAPH '11). ACM, New York, NY, USA, Article 1, 6 pages, 2011.

[16] L. Marg, L. C. Luri, E. O'Curra, and A. Mallett "Rating Evaluation Methods through Correlation", 1st Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE' 14), Reykjavik, Iceland, 26 May 2014.

[17] E. Meretzky, "Cavern of Doom", Pinnacle Books,New York, 1983,

[18] R. A. Montgomery, "Journey under the sea", Chooseco,2006

[19] E. Packard," Underground Kingdom", Bantam, 1983.

[20] M. E. Paschali, A. Ampatzoglou, A. Chatzigeorgiou, and I. Stamelos, "Non-functional requirements that influence gaming experience: A survey on gamers satisfaction factors", 18th Academic MindTREK Conference (MindTREK' 15), ACM, 4 – 6 November 2014, Tampere, Finland

[21] M. E. Paschali, N. Bafatakis, A. Ampatzoglou, A. Chatzigeorgiou, and I. Stamelos, "Tool-assisted Game Scenario Representation through Flow Charts", *13th International Conference on the Evaluation of Novel Approaches to Software Engineering (ENASE '18),* Madeira, Portugal, 23-24 March 2018.

[22] F. Peinado, P. Gervás, "Evaluation of automatic generation of basic stories, In New Generation Computing", Volume 24, Issue 3 , pp 289-302, 2006

[23] S. L. Pfleeger and B. Kitchenham, "Principles of Survey Research Part 1: Turning lemons into lemonade", Special Interest Group on Software, ACM, 26 (6), pp. 16-18, November 2001

[24] D. Pinelle, N. Wong and T. Stach, "Using genres to customize usability evaluations of video games", Proceedings of the 2008 Conference on Future Play: Research, Play, Share (Future Play'08), Association for Computing Machinery, pp. 129-136, Toronto, Ontario, Canada, 3 – 5 November 2008.

[25] P. Piselli, M. Claypool and J. Doyle, "Relating cognitive models of computer games to user evaluations of entertainment", Proceedings of the 4th International Conference on Foundations of Digital Games (FDG'09), Association for Computing Machinery, pp. 153-160, Orlando, Florida, USA, 26 – 30 April 2009.

[26] B. Rolfe, C. Jones and H. Wallace, "Designing dramatic play: story and game structure. In Proceedings of the 24th BCS Interaction Specialist Group Conference (BCS '10)." British Computer Society, Swinton, UK, UK, 448-452, 2010.

[27] R. B. Rookhuiszen and M. Theune, "Playful vs. serious instruction giving in a 3D game environment", Entertainment Computing, Elsevier, 1 (2), pp. 95-104, April 2009

[28] C. Roth, C. Klimmt, I. Vermeulen, P. Vordere, "The Experience of Interactive Storytelling: Comparing Fahrenheit with Façade", 10th International Conference, ICEC 2011, Vancouver, Canada, pp 13-21, October 5-8 2011

[29] C. Roth, P. Vorderer, C. Klimmt, "The Motivational Appeal of Interactive Storytelling: Towards a Dimensional Model of the User Experience", Second Joint International Conference on Interactive Digital Storytelling, ICIDS 2009, Guimarães, Portugal, Proceedings, pp. 38-43, 2009

[30] R. Rouse, "Game Design Theory and Practice (2nd ed.)", Wordware Publishing Inc., 44 Plano, TX, USA, pp 293-303, 2000.

[31] P. Runeson, M. Host, A. Rainer, and B. Regnell, "Case Study Research in Software Engineering: Guidelines and Examples", John Wiley & Sons, 2012.

[32] H. Schoenau-Fog, "Hooked! – Evaluating Engagement as Continuation Desire in Interactive Narratives", Fourth International Conference on Interactive Digital Storytelling (ICIDS 2011), Vancouver, Canada, pp 219-230, 2011

[33] P. Sweetser and P. Wyeth, "GameFlow: a model for evaluating player enjoyment in games", Computer and Entertainment. 3, pp.3-3, 3 (July 2005).

[34] C. Verbrugge, "A Structure for Modern Computer Narratives", Springer Berlin Heidelberg 2883, pp. 308-325,2003

[35] I. Vermeulen, C. Roth, P. Vorderer, C.Klimmt, "Measuring User Responses to Interactive Stories: Towards a Standardized Assessment Tool", Third Joint Conference on Interactive Digital Storytelling, ICIDS 2010, Edinburgh, UK, pp 38-43, November 1-3 2010

[36] S. Ware, M. Young, B. Harrison, D. Roberts, "Four Quantitative Metrics Describing Narrative Conflict", 5th International Conference, ICIDS 2012, San Sebastian, Spain, pp 18-29, November 12-15 2012.

[37] A. F. Wattimena, R.E. Kooij, J.M. van Vugt, O.K. Ahmed, "Predicting the perceived quality of a First-Person Shooter: The Quake IV G-model", Proceedings of 5th ACM SIGCOMM workshop on Network and System Support of Games, Singapore, October 2006.

[38] D. Weibel, B. Wissmath, S. Habegger, Y. Steiner and R. Groner, "Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment", Computers in Human Behavior, Elsevier, 24 (5), pp. 2274-2291, September 2008.

[39] C. Wohlin, M. Host, P. Runeson, M. Ohlsson, B. Regnell, and A. Wesslen, "Experimentation in software engineering: an introduction", Kluwer Academic Publishers, 2000