NExPlay
Playability Assessment for Non-experts Evaluators

Felipe Sonntag Manzoni, Bruna Moraes Ferreira and Tayana Uchoa Conte
Instituto de Computação, Universidade Federal do Amazonas, Av. General Rodrigo Octavio Jordão Ramos, 1200 Coroado I, Manaus, Amazonas, Brazil

Keywords: Playability, Heuristic Sets, Game Evaluation.

Abstract: Playability is related to the overall quality of the gameplay in a video game. It's important to evaluate the Playability to support games so they can fulfill the expectations of every player. However, given the high demand of game producers, limited budgets and deadlines, this type of evaluation is not always conducted, affecting the experience of players. One possible solution to evaluate Playability is through heuristic evaluations. Some researchers have conducted studies to develop heuristic sets that can satisfy the existing variety of games. However, given its broad comprehensiveness, those heuristic sets are too extensive and this may affect the feasibility of the assessment. Therefore, this work proposes NExPlay (Non-Expert Playability) Heuristic Set, which has the objective of minimizing the time and cost needed for the Playability assessment. Also, we aim to formulate a set that can be used by non-experts in Playability. The proposed heuristic set is assessed using a controlled experiment aimed at measuring the efficiency and effectiveness of our set, in comparison to another heuristic set. The results indicated that NExPlay identified problems with a more objective description. Furthermore, participants were able to understand the description of the heuristics presented in the NExPlay set more easily.

1 INTRODUCTION

The use of digital games, as well as their production, grows each year (Politowski et al., 2016). As a result of this growth, more and more game development companies are reducing their deadlines as well as their budgets for each game. These reduced deadlines and budgets causes a decrease in the evaluation of Playability, an important quality criterion for games (Aleem et al., 2016).

The main goal of Playability is to evaluate the existing interactions and relationships between the game itself and its design (Nacke et al., 2009). Through Playability assessment it is possible to evaluate game aspects going beyond the usual Usability evaluations of interfaces. It is possible to evaluate more specific game aspects, such as: mechanics, usability, history and game play (Desurvire et al., 2004). By considering these aspects, one can develop more attractive games for their target audience. However, many released games do not go through a testing plan or appropriate evaluation phases (Aleem et al., 2016). This may be related to the fact that some development companies do not have the availability of time and budget to carry out these evaluations at each stage of development. In addition, there is a lack of specialists in the context of Playability assessment, which makes its evaluation difficult. In this scenario, a more feasible alternative would be the use of heuristics to perform Playability evaluations. Heuristics are a simple way to help the inspector identifying key ideas needed for evaluations (Borys and Laskowski, 2014).

Some techniques for Playability evaluation using heuristics have been proposed (Desurvire et al., 2004; Korhonen and Koivisto, 2006; Desurvire and Wiberg, 2009; Barcelos et al., 2011). These are called heuristic sets which have the ability to evaluate softwares in a simple and fast way. Moreover, heuristic sets are usually developed to be specific for the context of its use, thus, reported problems always are related to the context stated in the heuristic set (Nielsen and Molich, 1990).

However, one of the problems with these heuristic sets is that game inspectors need to have previous experience with both the heuristic sets and the playability concepts, since heuristics are described in technical language (Desurvire and Wiberg, 2009). In

Sonntag Manzoni, F., Moraes Ferreira, B. and Uchoa Conte, T.
NExPlay.
DOI: 10.5220/0006695604510462
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 451-462
ISBN: 978-989-758-298-1

addition, most of these heuristic sets have too many heuristics becoming too extensive and requiring a lot of time for games evaluation. This makes heuristic sets too difficult to remember and memorize during digital game inspections (Korhonen et al., 2009). Finally, the heuristics usually proposed by these sets have a low level of specificity (Barcelos et al., 2011) and the language used in some heuristics is not easy to understand, making it more difficult for inexperienced inspectors to understand them (Korhonen and Koivisto, 2006).

In this context, we propose the NExPlay heuristic set, aiming at performing the evaluation of digital games by both specialist and non-expert inspectors. Non-expert inspectors are those who have low knowledge in game evaluations using heuristics. In addition, a small heuristics set is desired with heuristics that can be easily understood so that the inspectors can recall these heuristics during the evaluation process. This would help the inspectors recall which features should be evaluated in the game in subsequent evaluations, improving the productivity of the evaluation since they already know which heuristics can be used for which problem (Desurvire et al., 2004; Desurvire and Wiberg, 2009).

Considering the above, our goal is that the evaluations become faster in order to be possible to apply them in a short time. Finally, a reduction in costs is expected because inspections can be carried out by non-specialist inspectors.

To verify the feasibility of the NExPlay set, we conducted a comparative study between such set and the set proposed by Barcelos et al. (2011) to verify the efficiency, effectiveness and perception of the inspectors with regards to the use of these sets.

This paper is organized as follows: Section 2 presents related works to this research. Section 3 presents the construction of this heuristic set (NExPlay). Section 4 presents an empirical comparative study between the heuristic set purposed in this research and a heuristic set found in literature, which aims to analyze the effects of knowledge in the evaluations as well as its experimental process. Section 5 presents the quantitative and qualitative results of the comparative study. Then, Section 6 discusses the threats to the validity of the study. Finally, Section 7 presents the conclusion of this research and possible future work.

2 BACKGROUND

2.1 Playability

Nacke et al. (2009) indicated the main differences between Playability and Player Experience, proposing a more complex and accurate concept for Playability itself. According to its presented concept, the main goal of Playability is to evaluate the existing interactions and relationships between the game itself and its design. It mainly relates whether the information desired for an artifact conforms to its general design and how it presents itself.

The Player Experience is described as the relationship between the game itself and its user. That is, we are concerned with evaluating the interactions that users perform within the game and whether these activities are in accordance with what users expect (Nacke et al., 2009).

It is important to evaluate the Playability of digital games to identify interaction and design issues during their development. These problems can improve the overall quality of games, making them more accepted in the market by users (Sánchez et al., 2012).

2.2 Related Work

There are several researches in the field of development of heuristic sets for the evaluation of Playability in literature (Desurvire et al., 2004; Pinelle et al., 2008; Desurvire and Wiberg, 2009; Barcelos et al., 2011). These studies aim to produce standardized sets for all types of Playability evaluations, regardless of the analyzed artifact. However, in most cases, heuristic sets are too large for quick game reviews, and they may not cover all types of games. Furthermore, such a generalization approach of heuristics misleads evaluators so that they evaluate general rather than specific aspects of an artifact. This can be observed in the results described by Desurvire et al. (2004), where subjects described this difficulty with the use of very large and general sets.

Desurvire et al. (2004) developed a heuristic set called HEP (Heuristic Evaluation for Playability) containing 43 heuristics. During their analysis, their evaluators identified problems with its size as it became too long to be memorized. In addition, that set of heuristics (HEP) defines the categories of Game Play, Game Story, Mechanics and Usability in order to distribute the information between these aspects. These aspects are intended to guide the inspector during game evaluation so that the heuristics can be found more easily.

Desurvire and Wiberg (2008) describe, in a comprehensive way, the construction of a technique composed of guidelines with the goal of improving the initial experience of users through well-structured tutorials.

Desurvire and Wiberg (2009), in an evolution of the HEP heuristic set (Desurvire et al., 2004), developed the PLAY heuristic set. After this evolution, the set has 50 heuristics subdivided into several categories. However, the large number of heuristics also caused that the inspectors had difficulty to remember them and to find specific heuristics among the large number of aspects.

Korhonen and Koivisto (2006) proposed a set of heuristics with 29 heuristics in the context of mobile games (Mobile heuristics). The authors identified that, for the mobile context, there are specific problems for the mobility requirements of the system that must be taken into account.

Korhonen et al. (2009) performed comparative experiments between the heuristic set defined in their previous work (Korhonen and Koivisto, 2006) and a heuristic set found in literature. The authors discovered the strengths and weaknesses of the sets in order to identify patterns of production within the sets existing in the literature.

Heuristic sets are derived from the use and adaptation of sets from the literature (Desurvire et al., 2004). In addition, heuristics can be obtained through production guides and best practices, especially when we talk about game design. Heuristic sets have different ways to be applied to the real world, and they can be adapted to different situations. In most cases, heuristics sets are too long and not specific enough to be applied in specific game types (for example: Fight, Racing, Simulation).

In addition, we noticed that most of the sets found in the literature have a very similar structure. This structure considers a Game Play category, a Game Mechanics category and a Usability related to the game design category. However, this last category does not use the common concept of Usability applied, for example, by the heuristics defined by Nielsen and Molich (1990).

2.3 Heuristic Set Proposed by Barcelos et al. (2011)

Barcelos et al. (2011) defined a set of 18 heuristics. This set was based on other existing sets, seeking to decrease their size. Authors expected to see differences between reduced and large sets, consequently affecting the evaluation. In their evaluation, two similar games were used with regards to context and type. However, it has not been proven that the total number of heuristics has any effect on the evaluations. In addition, there is no separation into categories, as usually done in the literature. This heuristic set was chosen for the comparative study because it has a reduced number of heuristics, similar to our proposal. Its main difference is that it has no division into categories, which may help to study the effects caused by such categorization. The set used for comparison is available in Table 1.

Table 1. Heuristics proposed by Barcelos et al. (2011).

N°	Heuristics
H1	Controls should be clear, customizable and physically comfortable; Their response actions must be immediate.
H2	The player must be able to customize the audio and video of the game according to his/her needs.
H3	The player must be able to easily obtain information about everything in the surroundings.
H4	The game should allow the player to develop skills that will be needed in the future.
H5	The player must find a tutorial and familiarization with the game.
H6	The player must easily understand all visual representations.
H7	The player must be able to save the current state to resume the game later.
H8	Layouts and menus should be intuitive and organized so that the player can keep his focus on the game.
H9	The story must be rich and engaging, creating a bond with the player and his universe.
H10	The graphics and soundtrack should arouse the player's interest.
H11	Digital actors and the game world should look realistic and consistent.
H12	The main objective of the game must be presented to the player from the beginning.
H13	The game should propose secondary and smaller goals, parallel to the main objective.
H14	The game must have several challenges and allow different strategies.
H15	The pace of play should take into account fatigue and maintenance of attention levels.
H16	The challenge of the game can be adjusted according to the ability of the player.
H17	The player must be rewarded for his achievements clearly and immediately.
H18	The artificial intelligence must represent unexpected challenges and surprises for the player.

3 THE PROPOSED HEURISTIC SET: NExPlay

As mentioned above, most heuristic sets proposed in the literature are extensive and do not consider the language used to describe the heuristics, making them difficult to be understood by inspectors. To minimize these limitations, the NExPlay set was proposed. The goal of this set is to reduce the number of heuristics presented by other studies in the literature (Desurvire et al., 2004; Korhonen and Koivisto, 2006; Desurvire and Wiberg, 2009; Barcelos et al., 2011). In addition, it is expected that all types of inspectors can understand the heuristics even if they aren't experts in games or Playability. In order to meet such goal, the heuristic sets presented in Section 2 were identified and we analyzed the structure of these sets and their impact on their evaluation. We evaluated the differences between each approach and the results of the related experiments.

In this analysis, the heuristics were selected and categorized. For such categorization, we considered the three most commonly used categories in the evaluation of Playability: Game Play, Usability and Mechanics. During the selection, the heuristics were tested in common game scenarios identifying their importance during an evaluation. After the categorization and selection, it was necessary to minimize the number of heuristics in order to optimize the set by combining the heuristics. At the same time, new heuristics were formulated if it was necessary to cover all aspects of the evaluation. In addition, interviews with experienced players were conducted, discussing their experiences with various games. The essential points from each interviewee about what needed to exist in good quality games were noted and later analyzed.

Below, we present the categories that divide the set and the heuristics that compose each of these categories. All the heuristics presented below were formulated using 3 heuristic sets found in the literature: Korhonen and Koivisto (2006); Desurvire et al. (2004) and Desurvire and Wiberg (2009). Different heuristics, proposed by different authors, were merged and reformulated to minimize the total number of heuristics. As specified, the heuristics presented tend to address all the necessary aspects for the assessment of Playability in digital games.

3.1 Game Play

Game Play is the definition for the evaluation of what has been modeled and designed for that game in relation to the interactions of users with the game mechanics and with other players (Bjork and Holopainen, 2005).

More specifically, when we talk about Game Play, we want to evaluate how a game presents the primary and secondary goals to its players. Or, for example, how the game introduces the story of the game while accomplishing such goals as well as its tutorial. Also, it assesses how the game helps users solve these obstacles (i.e. we evaluate how users perform these interactions and how the game presents these interactions to the users) (Bjork and Holopainen, 2005). Table 2 presents the heuristics defined in the Game Play category.

Table 2. Heuristics for the Game Play category.

Game Play	
G1	Activities developed during the game are varied in order to reduce fatigue while also being balanced with the difficulty of the game.
G2	The game is balanced in order to apply pressure on the player without frustrating him/her. As a result, challenges are positive experiences for the player and keep him/her interested.
G3	The player should not be penalized repetitively for the same mistake or lose any object that was obtained through great effort.
G4	The artificial intelligence should present challenges and unexpected surprises for players (independent of their level) as well as keeping up with their learning.
G5	The skills needed to achieve a current or future goal are known and taught, and for such learned skills, the game offers a clear and immediate reward.
G6	The game has several secondary and optional goals which complement one or more main goals. Each of these goals must be achievable in different ways.

3.2 Usability

Regarding the concept of usability in the evaluation of Playability in digital games, we must highlight that the traditional concept should not be used. Therefore, we should define Usability for Playability so that it can be adapted for the desired needs of this concept. Initially, the Usability concept is commonly defined as the concern regarding the strict use of a product (effectiveness, efficiency and satisfaction in a particular context of use) (Bevan, 2001). That is, to facilitate the interaction between the users and their products within their context.

However, if we used this definition for the evaluation of Playability we would face the following inconsistency: the difficulty that a user faces using a game is intended, since challenges must be presented

to a player. Therefore, we cannot evaluate every game with the intention of reducing all the difficulties faced by the players. As a result, we want to use Usability with the intention of evaluating Game Design. And, in this context, the desired Game Design refers, in general, to the interfaces of communication and interaction with the user. That is, we want to improve the user's understanding of the information that is available to him/her. The information passed to the user must be correct, understandable and easy to locate. We can also mention the concern with the structuring and formulation of the tutorials that should be made present to the players. Table 3 presents the heuristics defined in the Usability category.

Table 3. Heuristics for the Usability category.

Usability	
U1	Every time the player starts a game, (s)he should have enough information to start playing, that is, there is no need to constantly access manuals, documentation or the tutorial.
U2	The player can easily get information about his surroundings plus information about the current state of the game, scores and other information in an obvious and simple way.
U3	The game presents feedback to the player appropriately and consistently, immediately and challenging with regards to the player's actions.
U4	Layouts and menus should be intuitive and organized so that the player can stay focused on the game as well as being able to avoid unintentional mistakes by the player.
U5	The player must be able to easily interrupt the game at any time, in addition to being able to save the current state to resume the game later.
U6	The game must present a well-structured initial tutorial to familiarize the player. Also, the challenge of the game can be adjusted according to the player's abilities.
U7	The player should easily understand all visual representations of the game.
U8	The tutorial should teach the player the basic functionalities and mechanics without being intrusive or creating a type of Game Manual during its extension.

3.3 Mechanics

The last category defined in our set of heuristics is Mechanics. The mechanics of a digital game are related to the concepts of how the controls are mapped according to the applied type of game. In addition, we can include concepts that refer to game settings, which directly interfere with how games behave.

We can still include aspects regarding the help that games should provide players so that they do not feel trapped in the game. This can be presented through different mechanics within the game environment itself. For example, a game could provide tips and tutorials that introduce the necessary actions to the players. Table 4 presents the heuristics defined in the Mechanics category.

Table 4. Heuristics for the Mechanics category.

Mechanics	
M1	The controls should be clear, customizable, and physically comfortable, and their response actions are immediate.
M2	The controls should be simple enough so that learning is fast, however, these should be expandable to more skillful players.
M3	Players should receive help from the context in which they find themselves, so that they are not tied to the point of needing a manual. However, such help should not facilitate too much the tasks that are necessary for a phase.
M4	The players must have the ability to modify the game settings. The way to change the settings must be simple and satisfactory.
M5	Terminologies used for all objects / functions / others can be understood by users.

4 EMPIRICAL STUDY

To verify the feasibility of NExPlay, we carried out a comparative study with the set proposed by Barcelos et al. (2011). This study aims to assess whether the developed set could evaluate a random game in a simple way, having both experts or novice inspectors in the field of games or inspections. We used this empirical study to evaluate if the set is constructed in order to support the evaluators at the time of the inspection of a game using an evaluation based on heuristics. We also wanted to know if the set covers all the necessary aspects so that the problems can be identified and associated with a heuristic. To do so, we evaluated the effectiveness and efficiency of the two sets and the perception of the subjects regarding their use.

24 undergraduates and graduate students in Computer Science participated in this empirical study. All subjects had prior knowledge of the concepts on Human-Computer Interaction and Software Engineering. The subjects had different levels of knowledge in digital games and software inspections. The empirical study was conducted to assess whether NExPlay is feasible for evaluating the playability of a game with expert inspectors or not.

To carry out the study, we used the two different sets of heuristics so that half of the subjects used the NExPlay set and the other half used the set proposed by Barcelos et al. (2011). The main difference between the two sets if that the set proposed by Barcelos et al. (2011) has no categorization of its heuristics, while NExPlay does. In this way, it can be assessed whether the division into categories causes some effect on the result or not. To guarantee that both groups have subjects with the same level of knowledge in inspection and games, we balanced the subjects.

To balance the subjects, all the evaluators answered a characterization questionnaire prior to the study. The characterization questionnaire included questions regarding the overall experience of the subjects in Software Inspection. For each question, 4 different levels were defined: no experience, low, medium and high. These levels were defined according to the academic knowledge and the industry application of these characteristics by the evaluators.

In addition to questioning the experience of software inspection, we also asked the subjects regarding their previous experiences with games in general. To do so, we defined 6 questions of experience in this field that varied according to the number of times per week that they used games on a daily basis.

The subjects were assigned randomly to both groups (principle of random design), respecting the balance of experience in inspection and games. At the end of the balancing, each of the two groups had 12 subjects (24 subjects in total).

4.1 Experimental Process

Initially, we selected a game to be evaluated by all subjects. In order to carry out this evaluation, we selected a casual game type: Leap of Cat, which presents a simple theme and story, accessible only through its description in the app store itself. Leap of Cat has been chosen as it is freely available for the Android platform through the Google Play store on any compatible device. In addition, it presents several problems related to Playability even though it has a positive evaluation by users.

Before conducting the experiment, it was necessary to introduce the fundamental concepts of Playability to all subjects. For this, we carried out a training with all subjects. The training consisted of presenting the core concepts of Playability so that subjects could have the same knowledge base.

After this training, the subjects were instructed to use the game and the heuristic set to find problems in the application. The heuristic sets, together with the problem specification and classification table, were sent to the individual e-mail of each subject so that none of them could know what each subject would be using. The subjects could perform the evaluation of the game at any time until a certain date. Each subject should send his/her results separately to the inspection organizers.

Each evaluator performed the inspection individually, writing their respective times in the table of discrepancies. This table with the description of the identified discrepancies was sent to the researchers. After receiving all tables, a list of discrepancies without duplicates was created. This list was analyzed by two experts who classified all discrepancies in problems or false-positive. Table 5 shows the final list containing the number of unique defects and false-positive from this experiment.

The treatments for the independent variable are the two employed heuristic sets and the dependent variables are the efficiency and effectiveness of the sets. The efficiency was calculated for each subject as the ratio between the number of defects and the time spent evaluating the artifact. The effectiveness was calculated for each subject as the ratio between the number of defects found and the total number of defects (known) for the artifact. The experiment was performed to test the following hypotheses (null and alternative respectively):

H01: There is no difference in terms of efficiency when using the NExPlay heuristic set and the set defined by Barcelos et al. (2011).

HA1: There is a difference in terms of efficiency when using the NExPlay heuristic set and the set defined by Barcelos et al. (2011).

H02: There is no difference in terms of effectiveness when using the NExPlay heuristic set and the set defined by Barcelos et al. (2011).

HA2: There is a difference in terms of effectiveness when using the NExPlay heuristic set and the set defined by Barcelos et al. (2011).

5 RESULTS

To facilitate reporting the results, we will refer to the heuristics sets as: Set 1 for the NExPlay set developed by the authors of this paper, and Set 2 for the set presented by Barcelos et al (2011). Likewise, the first group, which used Set 1, will be called Group 1, and the second group that used Set 2 will be called Group 2. The results are presented below.

5.1 Quantitative Results

Overall, 49 unique problems were identified in the evaluated game considering all problems from Set 1 and Set 2. Set 1 found 21 problems among these unique problems and Set 2 found 38 problems among these unique problems. Regarding the knowledge levels of the subjects, we have presented the individual knowledge levels of each of the subjects, as shown in Table 5.

Table 6 presents the calculated means for the effectiveness and efficiency indicators for each of the sets. We performed statistical tests for the efficiency and effectiveness of the sets.

Table 6. Means for effectiveness and efficiency by set.

Sets	TP	M	AE (%)	TT (h)	EF
Set 1	21	6,16	42,86	9,26	2,27
Set 2	38	8,41	77,55	10,17	3,74
Legend: TP – Total Problems; M – Mean (Problems); AE – Average Effectiveness; TT – Total Time; EF – Efficiency (Problems/Hour)					

In order to decide which statistical tests would be necessary for comparing the efficiency and effectiveness, normality tests were performed for each of the groups. To compare the samples related to efficiency, we used the t-student test with a confidence level of 0.05. We selected the t-student test based on the results of the Shapiro-Wilk normality test, which indicated that the variables (efficiency values) were normally distributed. To compare the samples related to effectiveness, we used the Mann-Whitney test with a confidence level of 0.05. We selected the Mann-Whitney test based on the results of the Shapiro-Wilk normality test, which indicated that the variables (effectiveness values) were not normally distributed. We performed the statistical analysis using the SPSS tool. The normality tests for the two samples are presented in Table 7.

To compare the efficiency of the two sets, we used a boxplot analysis and the parametric t-student test. To compare the effectiveness of the two sets, we used a boxplot analysis and the non-parametric Mann-Whitney test. Figure 1 shows the boxplots comparing the distribution of effectiveness and efficiency per set.

Table 5. Results for each inspector.

N°	Knowledge in Inspections	Knowledge in Games	Number of discrepancies	Total false-positive	Total Problems	Time (h)	Efficiency (Problems/ Hour)	Total Defects
01	Low	Low	4	0	4	0,67	6,00	**NExPlay = 74**
02	Low	Low	4	1	3	0,83	3,60	
03	Medium	None	9	3	6	1,97	3,05	
04	High	Low	11	5	6	1,25	4,80	
05	Low	Medium	6	0	6	0,53	11,25	
06	Low	Low	7	1	6	1,17	5,14	
07	High	None	3	1	2	1,05	1,90	
08	Low	Low	8	1	7	0,63	11,05	
09	Medium	Low	14	4	10	1,00	10,00	
10	Low	High	18	3	15	0,92	16,36	
11	Medium	Low	5	0	5	1,00	5,00	
12	Medium	High	6	2	4	0,42	9,60	
13	High	High	20	5	15	1,50	10,00	**Barcelos et al. (2011) = 101**
14	Medium	Low	5	1	4	1,18	3,38	
15	High	Medium	18	3	15	0,92	16,36	
16	Low	Low	13	1	12	1,85	6,49	
17	Low	None	9	1	8	0,60	13,33	
18	None	None	4	1	3	0,63	4,74	
19	Low	Low	3	0	3	0,57	5,29	
20	Low	Low	10	0	10	0,78	12,77	
21	Medium	Low	9	0	9	0,92	9,82	
22	None	Low	5	1	4	1,35	2,96	
23	Low	High	7	1	6	1,23	4,86	
24	High	None	12	0	12	1,42	8,47	

Table 7. Normality test – Efficiency and Effectiveness.

	Shapiro-Wilk	
Efficiency	Group	Significance
	Set 1	0,283
	Set 2	0,407
	Shapiro-Wilk	
Effectiveness	Group	Significance
	Set 1	0,033
	Set 2	0,208

When comparing the two sets with regards to efficiency using the t-Test, no significant statistical difference was found between the two groups ($p = 0.618$). These results suggest that Set 1 and Set 2 provided similar efficiency when used to inspect digital games. The same analysis was applied to verify if there was a significant difference regards to the effectiveness indicator of the two sets in the detection of Playability problems.

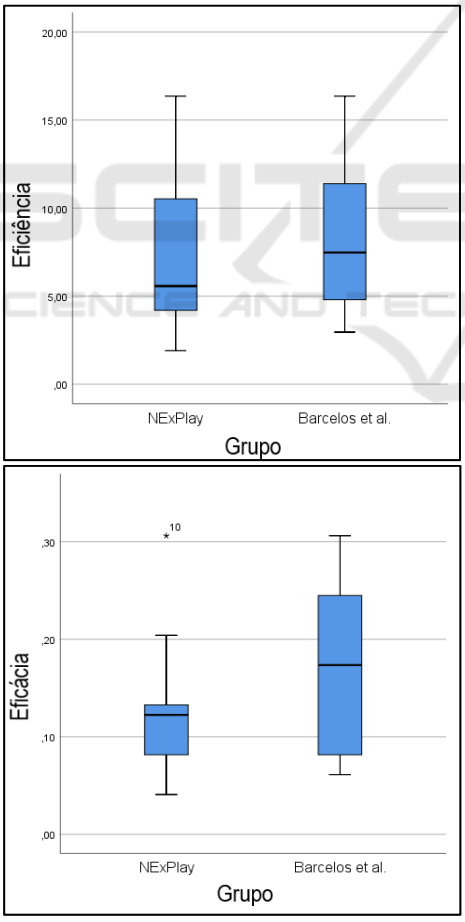

Figure 1. Box Plot Graphic - Effectiveness and Efficiency.

The Mann-Whitney tests showed that the two sets, regards to effectiveness, did not obtain significant statistical differences between the groups ($p = 0.266$). These results suggest that Set 1 and Set 2 provided similar effectiveness when used to evaluate digital games. According to these quantitative results, it is not possible to reject the null hypotheses attributed to this experiment.

5.2 Qualitative Results

In order to understand the subjects' perceptions regarding the applied heuristic sets, questionnaires were completed after the inspection. The questionnaires were provided online to the subjects. In addition to the questionnaires after using the sets, interviews were conducted with some of the inspectors to identify more details about their difficulties and perceptions about the applied sets. The results are described below.

5.2.1 Heuristic Specificity

With regards to Set 1, some evaluators pointed out that, due to the design decision of the heuristics, the union of more than one problem in the same sentence (heuristic) affected its judgment at the moment of indicating a heuristic. This can be observed by the comment made by the evaluator 11:

"(...), the heuristics have many aspects being considered in the same heuristic. For example, I sometimes agreed with the first part of the heuristic, but the second part did not fit. So, I would rule it out." – Evaluator 11

That is, according to the evaluators, the use of sentences with more than one problem in each heuristic being described, affected their understanding of the whole problem being described and whether the problem should be identified by such heuristic. This could be corrected, breaking the heuristics that have such characteristic in several heuristics. Evaluator 09 describes the same problem:

"...it has heuristics that make you confused as you answer. Because one part of it is true and the other is not. And when we answer, we do not know if we have to describe in a general way or answer only with regards to a part of the heuristic" – Evaluator 09.

This problem can also be observed in Set 2 through the comments made by evaluator 16:

"...the heuristics that have two levels of verification, such as A10 – 'The graphics and the soundtrack should increase the interest of the player', could be separated to be better identified by the verifier ..." – Evaluator 16.

Evaluator 9 indicated that he wanted the presented heuristics to be more dispersed in a larger number of heuristics. That is, the problems indicated through each of the heuristics were simpler and smaller to be evaluated. This can be observed through:

"The heuristics should be divided... My suggestion would be to review the heuristics and make them more unique..." – Evaluator 9.

5.2.2 Categorization of the Heuristic Set

Evaluator 08 (who used the NExPlay set) pointed out that, the categorization of the heuristics in Game Play, Usability and Mechanics was unnecessary. In addition, the same evaluator pointed out that, in these categories, there are heuristics that are similar.

"Some heuristics (G4 e U1) evaluate the same thing. I think that Game Play, Usability and Mechanics could be unified (once this categorization would not be understood by users anyway) asking all the heuristics in sequence, caring out the repeated heuristics which evaluate the same thing (G4, U1 and U8)" – Evaluator 08.

On the other hand, regarding the categorization, Evaluator 02 pointed out that, with regards to Set 1, the categorization of the heuristics in these 3 aspects helped during the evaluation. For the most part, the evaluators did not point out exactly at which points, or heuristics, the set was similar, as shown in the comments by evaluator 11:

"The technique has many heuristics, some very similar to others..." – Evaluator 11.

From the analysis of these results we can see that the use of this type of categorization did not affect the performance in the inspection process. Despite this, Evaluator 10 found it positive to have a separation into categories:

"As positive points, I highlight the small number of heuristics and their categorization (...)" - Evaluator 10.

5.2.3 Number of Heuristics

As mentioned before, it was indicated that the number of heuristics was very large. A possible solution would be to build sets of specific heuristics for the types of game presented for evaluation. However, some evaluators specified that the set was small and broad:

"I found the set of heuristics with a good completeness and a small size, not becoming tiring. Also, they were not repetitive" – Evaluator 10.

Evaluator 6 pointed out that the presented heuristics could not cover some problems presented by the evaluated games:

"...some heuristics may not cover some problems presented by the game" – Evaluator 6.

However, it is not clear what these problems are that the set may not evaluate. Some evaluators explained that the heuristics could be more objective facilitating the understanding of the evaluators. This can be observed specifically in the sentence described by evaluator 3:

"...these concepts are very extensive, perhaps a more objective phrase about the heuristics, would facilitate the interpretation of the evaluator" – Evaluator 3.

Based on these results, future studies should consider that the heuristics must be reformulated by reducing its extension and increasing its specificity.

5.2.4 Understanding of the Heuristics and Game Type Targeting

Regarding the repeatability and similarity between heuristics in Set 2 we have the comments made by evaluator 24:

"I found some similar items and it was confusing to understand the difference between them. For example, A14 and A18 items seemed to address the same thing" – Evaluator 24.

Some evaluators have indicated that heuristics from Set 1 can be easily understood:

"I did not find any difficulties in understanding the heuristics" – Evaluator 2.

"The language used in the technique is simple, I had no difficulty in understanding, even though I did not have much experience with game evaluation" – Evaluator 1.

With regards to the comprehension of the heuristics on Set 2, some difficulties were reported regarding their understanding:

"I could not associate / evaluate this heuristic: 'The pace of the game must take into account fatigue and maintenance of attention levels'. I was not sure if what I had to evaluate was the stressfulness to use the App. As it was not clear, I did not evaluate this heuristic" – Evaluator 20.

"The heuristic 'Digital actors and the game world should look realistic and consistent'. I do not know if the term realistic would be ideal (...). I had this doubt about of what would be realistic in the context of the evaluation" – Evaluator 20.

According to some inspectors, the difficulty in understanding occurred both because some words used were very specific to games, and because others were confusing for their evaluation. However, in some cases the evaluators did not specify why they encountered difficulties with these heuristics or

words. This can be observed by the comments of the evaluators 16, 22 and 24:

"Heuristic A15 is not very clear (...). I was hesitant to use the A03 or A05 heuristic for the game problem of having no help available" – Evaluator 16.

"Some terms seem confusing as: easy, ease, rich, challenges and strategies" – Evaluator 22.

"I found some similar items and it was confusing to understand the difference between them. For example, A14 and A18 items appear to address the same thing. On some items, there are some terms that leave the evaluator confused. For example, artificial intelligence. It was not clear if the artificial intelligence that the item refers to is the game" – Evaluator 24.

We can observe during the evaluation of the problems reported by the two groups, that the subjects who used Set 2 had difficulties in understanding the heuristics. On the other hand, no problems were reported regarding the understanding of the heuristics from Set 1. These are indications that the NExPlay set can be more easily understood, making it feasible for inspectors with no experience in Playability evaluation. Some inspectors pointed out that the heuristics could be targeted to different game types. That is, different heuristics should be created which could evaluate the different types of existing games, since the presented heuristics are focused on a general context. This can be observed in the problem described by evaluator 12:

"...Some heuristics are not compatible with all types of games. The issue of customization of controls (M1 and M2) and artificial intelligence (G4), for example, are not suitable for the evaluated game. This makes me wonder if I should report that the game does not allow this customization or if I should consider that the game is simple and does not have the possibility to incorporate these aspects..." – Evaluator 12.

Thus, this lack of categorization of the heuristics with regards to the evaluated game type affected the evaluation performed by the evaluators. The same observation was made by evaluator 23 in Set 2 as described below:

"The heuristics cover relevant game features in general, but specific types of games must not necessarily have all these characteristics, which makes the inspector report problems that do not apply to certain game types" – Evaluator 23.

5.2.5 Interviews

For a better understanding of the application of the heuristics and to make improvements in the NExPlay Set, 4 evaluators were selected for an interview. These evaluators were chosen due to their lowest problem rate, when compared to the other inspectors. Thus, interviews with evaluators 1, 2, 7 and 11 were performed. The interviews were conducted in order to know the reasons that led these evaluators to have such rates as well as their overall difficulties. In addition, we wanted to know some specific aspects of the set and how it was applied.

Evaluator's 1 and 11 identified that some heuristics would not be applied to the context of the evaluated game, because the game type was too simple to perform the inspection of all the aspects covered by the heuristics. Thus, we consider that this is a problem regarding the use of heuristics not directed to the type of games, but heuristics formulated to cover generalized contexts.

The evaluator also explained that using heuristics as a guide helped to avoid any game problem from being identified. That is, the evaluator read each of the heuristics and then found a way to test such heuristic in the game to find out if it identified a problem in the game or not. The evaluator himself indicated that he decided to carry out the inspection that way due to his knowledge.

Evaluator 2 pointed out that the leaping mechanics of the main character of the game presented difficulties so that it took considerable time to be learned. However, the evaluator did not indicate such problem during his game inspection. Therefore, we should verify if the heuristics presented to the evaluators, as well as their mechanics and usability, are oriented or are explicit enough to be identified. He also identified that one of the problems encountered during his inspection aroused directly from the reading of the heuristics. This problem was identified by heuristic G4, in which it described the lack of modification of the environment, difficulty and surprises in the game.

When asked about the use of a categorization for different sets of heuristics, Evaluator 2 also stated that the classification into distinct groups helped and was used by him during the game inspection process. He evaluated the game separately in terms of Game Play, Usability and Mechanics. Moreover, when he had difficulties with the classification of some problem, he classified this problem following the logic of the specified categories. Considering such information, we should further evaluate the use of these categories in other types of evaluations instead of Playability in order to identify their use outside this context.

As pointed out in his questionnaire, Evaluator 2 reinforced that there is a need to create heuristics related to the different game types, since, for the most

part, the presented heuristics were described in a generic way.

Evaluator 7 indicated that he did not have difficulties with the language used to describe the heuristics. However, the heuristics may not have a correct targeting since, during the interview, the evaluator indicated problems encountered by him with respect to the leaping mechanics, which did not appear in his description of the problems. The evaluator also pointed out that the heuristics of the Mechanics category had very long descriptions. Such descriptions affected the decision of the evaluator at the time of identifying a problem.

Evaluator 11 indicated that the set had repeated information regarding heuristics U9, U10 and U15. In these, the evaluator indicated that all were evaluating the same type of object, which concerned the visual information of the game. In this way, the evaluator indicated that during his inspection he had to evaluate the same aspect several times disrupting his concentration and development. The categorization of the heuristics did not affect the evaluator's inspection at all.

5.2.6 Description of the Identified Problems

In addition to collecting the subjects' perceptions regarding the use of the heuristic sets, an analysis of the descriptions of the defects found by the subjects was performed. All defects found by the two sets were analyzed by 3 specialists, to see if they could be understood. The analysis showed that the descriptions of the defects reported by Set 1 had a more objective description. That is, the description of the problems could be better understood with regards to the observed problem by the evaluator and its location in the evaluated artifact.

In some cases of defects founded by NExPlay, we observed that the description of the problem followed the description of the affected heuristic itself, while locating the problem and the moment when it happened. This fact was not observed in the defects described by the subjects who used the set proposed by Barcelos et al (2011).

6 THREATS TO VALIDITY

The threats related to this study were divided into four categories: internal validity, external validity, conclusion validity, and construct validity (Wohlin et al., 2012).

Internal Validity: There could be an effect caused by the training given if the training of Set 1 had a lower quality than the training given to Set 2 and vice versa. This threat was controlled guaranteeing that the same training on Playability was taught to all subjects using generic examples applied to games not related to the assessed game. Regarding the classification of knowledge, this was a self-classification carried out by the subjects in relation to their previous knowledge (Games and Inspection). With regards to time measurement, this could have a significant impact on the results, since we did not have control of how the subjects recorded such times and we cannot be sure if they recorded their correct and real times. In an attempt to control such a threat, the form delivered by the subjects had distinct fields for the start and end times of the game phase and the inspection phase.

External Validity: The “Leap of Cat” game used for the evaluation does not represent all types of existing games. The evaluators in this study were not practitioners from the gaming industry or Playability experts. However, this cannot be considered a real threat since the target audience was non-expert evaluators. Then the sample represents the population.

Conclusion Validity: The number of subjects is not ideal from a statistical point of view.

Construct Validity: For this type of threat, the definitions for the efficiency and effectiveness indicators were considered. Such indicators are commonly used in studies investigating the detection of defects. In addition, these indicators were used in the same way as in other studies (Korhonen et al., 2009; Valentim et al., 2015).

7 CONCLUSION AND FUTURE WORK

According to the results presented by the empirical study, we cannot infer any conclusion regarding the quantitative results of the statistical tests. However, from the analysis of the data regarding the subjects' perception regarding the use of the sets, it was possible to observe that some subjects liked the categorization of the heuristics. Nevertheless, this type of categorization was not a sufficient factor for the NExPlay set to perform better than the set proposed by Barcelos et al (2011). To verify such a difference, more studies should be carried out with a larger number of subjects.

In addition, some subjects agreed that the heuristics defined in the NExPlay set could be easily understood. However, we noticed that some

evaluators questioned the completeness of the set where more heuristics would be necessary. Moreover, some subjects had difficulties in understanding the heuristics of the set proposed by Barcelos et al (2011). Thus, taking into account the sample used, there are indications that NExPlay supports the inspection of games by inexperienced evaluators since the problem regarding the evaluators' understanding was not observed.

One limitation of NExPlay is that it does not cover all types of digital games. Consequently, improvements will be necessary to adapt the heuristic set to the different existing game types so that all of them can be evaluated.

As future work, a new version of the NExPlay heuristic set will be developed, focusing on the issues related to size, completeness, organization and redundancy of the presented heuristics. In addition, such set will be adapted to evaluate each game according to its type. After the improvements, new studies will be performed using NExPlay to evaluate different types of games. Furthermore, we intend to evaluate the inclusion of the acceptability concept in a new version of the heuristic set, in order to assess emotions and the power of actions.

ACKNOWLEDGEMENTS

We would like to thank the financial support granted by CNPq through process number 423149/2016-4, and CAPES through process number 175956/2013.

REFERENCES

Aleem, S., Capretz, L. F. and Ahmed, F., 2016. Game development software engineering process life cycle: a systematic review. *In Journal of Software Engineering Research and Development, vol. 4.* Springer Open.

Barcelos, T. S., Carvalho, T., Schimiguel, J. and Silveira, I. F., 2011. Comparative analysis of heuristics for digital game evaluation. *In IHC+CLIHC'11, 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction.* Brazilian Computer Society. (in Portuguese)

Bevan, N., 2001. International standards for HCI and usability. *In International Journal of Human-Computer Studies, vol. 55, issue 4.* Elsevier.

Bjork, S. and Holopainen, J., 2004. *Patterns in game design (game development series)*, Charles River Media. Rockland, USA, 1st edition.

Borys, M. and Laskowski, M., 2014. Expert vs Novice Evaluators - Comparison of Heuristic Evaluation Assessment. *In ICEIS'14, Proceedings of the 16th International Conference on Enterprise Information Systems*. SCITEPRESS.

Desurvire, H., Caplan, M. and Toth, J. A., 2004. Using heuristics to evaluate the playability of games. *In CHI'04 extended abstracts, Conference on Human Factors in Computing Systems.* ACM Press.

Desurvire, H. and Wiberg, C., 2008. Evaluating user experience and other lies in evaluating games. *In CHI'08, Conference on Human Factors in Computing Systems.* ACM Press.

Desurvire, H. and Wiberg, C., 2009. Game usability heuristics (PLAY) for evaluating and designing better games: The next iteration. *In OCSC'09, International Conference on Online Communities and Social Computing.* Springer.

Korhonen, H. and Koivisto, E., 2006. Playability heuristics for mobile games. *In MobileHCI'06, 8th Conference on Human-Computer Interaction with Mobile Devices and Services.* ACM Press.

Korhonen, H., Paavilainen, J. and Saarenpää, H., 2009. Expert review method in game evaluations: comparison of two playability heuristic sets. *In MindTrek'09, 13th International MindTrek Conference: Everyday life in the ubiquitous era.* ACM Press.

Nacke, L., Drachen, A., Kuikkaniemi, K., Niesenhaus, J., Korhonen, H., Hoogen, W., Poels, K., IJsselsteijn, W. and Kort, Y., 2009. Playability and player experience research. *In DiGRA'09, 2009 DiGRA International Conference: Breaking New Ground: Innovation in Games, Play, Practice and Theory.* DiGRA Digital Library.

Nielsen, J. and Molich, R., 1990. Heuristic evaluation of user interfaces. *In CHI1990, Conference on Human Factors in Computing Systems.* ACM Press.

Pinelle, D., Wong, N. and Stach, T., 2008. Heuristic evaluation for games: usability principles for videogames design. *In CHI'08, Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems.* ACM Press.

Politowski, C., de Vargas, D., Fontoura, L. and Foletto, A., 2016. Software Engineering Processes in Game Development: a Survey about Brazilian Developers. *In SBGames'16, XV Simpósio Brasileiro de Jogos e Entretenimento Digital.* SBC.

Sánchez, J., Vela, F., Simarro, F. and Padilla-Zea, N., 2012. Playability: Analysing user experience in video games. *In Journal of Behavior & Information Technology, vol. 31.* Taylor & Francis Online.

Valentim, N., Conte, T. and Maldonado, J., 2015. Evaluating an Inspection Technique for Use Case Specifications Quantitative and Qualitative Analysis. *In ICEIS'15, Proceedings of the 17th International Conference on Enterprise Information Systems.* SCITEPRESS.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, C., Regnell, B. and Wesslén, A., 2012. *Experimentation in Software Engineering,* Springer-Verlag Berlin Heidelberg. Berlin, 1st edition.