Heuristic usability evaluation on games: a modular approach

Rosa Yanez-Gomez[1] · Juan Luis Font[1] · Daniel Cascado-Caballero[1] · José-Luis Sevillano[1]

Abstract
Heuristic evaluation is the preferred method to assess usability in games when experts conduct this evaluation. Many heuristics guidelines have been proposed attending to specificities of games but they only focus on specific subsets of games or platforms. In fact, to date the most used guideline to evaluate games usability is still Nielsen's proposal, which is focused on generic software. As a result, most evaluations do not cover important aspects in games such as mobility, multiplayer interactions, enjoyability and playability, etc. To promote the usage of new heuristics adapted to different game and platform aspects we propose a modular approach based on the classification of existing game heuristics using metadata and a tool, MUSE (Meta-heUristics uSability Evaluation tool) for games, which allows a rebuild of heuristic guidelines based on metadata selection in order to obtain a customized list for every real evaluation case. The usage of these new rebuilt heuristic guidelines allows an explicit attendance to a wide range of usability aspects in games and a better detection of usability issues. We preliminarily evaluate MUSE with an analysis of two different games, using both the Nielsen's heuristics and the customized heuristic lists generated by our tool.

Keywords Heuristic evaluation · Games · Usability · Heuristic guidelines

1 Introduction

The research on specific usability evaluation methods for video games (henceforth games) is a challenging area of study, which remains an upward trend [58]. The two possible approaches to address the special characteristics of games are the adaptation of traditional methods and the proposal of new ones. Both are key given the need of evaluations, which can assess some degree of success precognition before the release of new games to the market, especially since these software projects are extremely costly [14].

[1] Department of Computer Technology and Architecture, ETS Ingeniería Informática, Universidad de Sevilla, Avda Reina Mercedes s/n, 41012 Seville, Spain

The term usability when applied to games is subject of analysis and constant reformulation in order to cope with the specificities of games, which differentiate them in many respects from productivity software [29, 58]. Since the usage of games is motivated by enjoyment and fun, the interaction with gamers should be especially careful, avoiding interruptions and obstacles. Thereby, interaction must be fluid since usability is strongly tied to "*the degree to which a player is able to learn, control and understand a game*" [48] Thus, while "*the goals of software productivity are to make the software interface easy to learn, use, and master*", the design goals for games are usually characterized as "*easy to learn, difficult to master*" [12, 40].

Definitely, the users' attitudes towards games are different from the ones towards productivity software applications. Whether within anarchic or rule-based gameplay —free-form paideia and rule-bound ludus in Callois terms [5]—, satisfaction and fun are important parts of the experience of use. Nonetheless, the inclusion of these aspects into the reformulation of usability definition for games is matter of controversy. While some authors as Pinelle, Wong and Stach [48] distinguish "*entertainment, engagement, and storyline*" considering them "*strongly tied to both artistic issues (e.g. voice acting, writing, music and artwork) and technical issues (graphic and audio quality, performance issues)*" but not related to game usability, others such as Papaloukas, Patriarcheas and Xenos [47] include these aspects into their definition of game usability: "*the degree to which a player is able to learn, control, understand, be intrigued and enjoy a game*". Indeed, Papaloukas, Patriarcheas and Xenos even consider that game usability and fun cannot be measured separately because "*fun is a prerequisite of usability*" [47].

More conservative definitions emphasize different aspects of usability while maintaining its traditional definition: *"the capability to be used by humans easily and effectively"* [2], *"quality in use"* [10] or *"the effectiveness, efficiency and satisfaction with which specified users can achieve goals in particular environments"* [26]. This last definition leads to the ISO definition of usability [23]: "*"The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"*. Clarifying nuances and adding some reformulations according to the particularities of games, this definition seems to be useful as a base framework for research [58]. However, while some authors as Frøkjaer, Hertzum and Horbæck [15] consider these components —effectiveness, efficiency and satisfaction— as separate and independent aspects, others as Federoff [14] declare that all three measures are not equally important or applicable and emphasize one aspect over the others: "*in the case of video game usability, effectiveness and efficiency are secondary considerations in relation to satisfaction*" moreover when "*a consumer may need to purchase or use other software to perform necessary tasks, but a game is bought on a voluntary basis purely for entertainment value*".

Definitely, satisfaction is a fundamental part of the gaming experience. Thus, *playability* is the term coined to refer to motivational factors such as enjoyment, engagement and fun when playing games [52]. New expert methods focused on its evaluation have been proposed in the literature [5, 7, 14, 31, 40]. To Lennart Nacke [43] even "*admitting that usability and playability may be dichotomous concepts, they build a high level game usability framework model*". However, motivational processes are less well understood than cognitive ones involved in human-computer interactions [39], which add special difficulties to perform analysis of this aspect.

Additionally, popularization of networked games added a social component to games. The term *social usability* [24] [16] —also "*social playability*" [24]— tries to cover issues emerged

from the communities formed through communication and cooperation between users and players who interact with the same product. Going deeper, Järvinen et al. [24] defined different aspects as *functional playability* —referred to control mechanisms and their relation to gameplay—, *structural playability* —related to the aesthetics of digital games and entertainment (rules, strategy, etc.)—, *audiovisual playability* added to the previously described *social playability*.

The explosion of redefinitions, and the fuzzy boundaries of terms, increases with the inclusion of the user experience term. User experience is an umbrella covering all the process of acquisition and maintenance of a product in a wider conception of usability, which apply to any interaction with the product beyond the main functionality [13].

Integral factors of user experience are the state of flow and immersion [22, 28]. The concept of flow was first introduced in Csikszentmihalyi in 1975 [9] and defines optimal experience as a specific state of psychic energy in one's consciousness. To Johnson and Wiles [25] flow is a state of concentration, deep enjoyment, and total absorption during an activity. According to Hassenzahl [18] flow is a positive experience caused by an optimal balance of challenges and skills in a goal-oriented environment. Sweetser and Wyeth [55] take the eight elements from Csikszentmihalyi's concept of flow [9] and map them onto computer games creating the GameFlow approach for flow. However, the GameFlow approach has been criticized by Cowley et al. [8, 28]. Currently a clear definition and grounded understanding of this term is still missing [28, 35]. According to Law et al., the main problem in evaluating is that user experience treats non-utilitarian aspects of interactions between humans and machines [28, 35].

Regarding usability the interest in flow and immersion is not only about fun measurement but also the overseeing of usability issues while gamers are immersed [6, 28]. Brown and Cairns [4] distinguished three phases of immersion: engagement, engrossment and total immersion. Engagement is based on the interest the players have in the game. When engrossed in a game, the player's emotions are directly affected by this. Total immersion comes when the player is completely immersed in the game and experiences absolute presence, a situation where only the game and the emotions it inspires matter. Larsen [34] states in his work that common game reviews are to a major extent based on the subjective evaluation of a game's user experience from the game reviewer's point of view so bias derived from immersion and flow should be taken into account.

Beyond semantic issues, the market is constantly innovating in game controls, devices and interaction models, which offer enjoyable challenges for players, and give rise to new genres. Moreover, the variability in knowledge, expertise or cultural environment of potential gamers should be managed to achieve inclusion and commercial success. The disparity between players goes even further in *serious games* —"games used for purposes other than mere entertainment" [54]— which are usually played by users with special interaction needs, such as children or elderly users with different cognitive characteristics and sensory impaired people, as in the treatment of a wide range of pathologies, which can either directly or collaterally affect the interaction scenario. All of these constraints contribute to the special relevance of usability assurance as an unavoidable requirement, which crosses the design process.

This paper will pay special attention to heuristic evaluation (henceforth HE) as a convenient and flexible method. It can cope with the semantic questions previously presented as well as with the special characteristics of games and its innovative nature. HE and these required adaptations are discussed in the next section. To overcome the disparity of terminology and

semantics of the different proposals in this area present in the literature, and the limitations the individual application of them, a tool is presented in Section 3. This tool, named MUSE (MetaheUristics uSability Evaluation tool), takes potentially useful heuristic guideline items from the literature and rebuilds them into a customized list according to a tag-based description of the usability evaluation plan, that is, the characteristics of the game to evaluate, the goals of the evaluation and the development status of the product. In Section 4, the analysis required to develop MUSE is presented in detail. Specific subsections attend to the questions arisen from selection criteria of the heuristics used as the core of MUSE, the processing and filtering of these heuristics and the final definition of metadata tags used to establish categorizations. In Section 5, a preliminary evaluation is presented in order to validate some of the design decisions. A pilot test is followed up by several usability evaluations using the tool from which to measure the performance of MUSE. Finally, Section 6 summarizes conclusions and points some future challenges in the evolution of the tool.

2 Heuristic evaluation methods on games

Among the most commonly used game evaluation methods, playtesting is the preferred academic method for evaluating functional prototypes or alpha-versions of new games [58]. Expert-based methods are not widely used although they are considered the more apt for early phases of development and first prototypes [44]. Expert-based methods do not provide information on how players will experience the game but can predict problems before the game is released [30]. Expert-based evaluations are cheaper and convenient in many cases, while playtesting is not only impossible to run without playable prototypes, but it is typically outsourced to dedicated companies, making these preliminary releases risky for companies since they can lead to public information leaks too early altering users expectations [37].

When evaluators choose expert-based methods, HE is the most frequently selected methodology [58]. HE is well-known, low-cost and fast [38] and it is estimated that it is capable to discover up to 80% of major usability issues of an interface [46]. This convenience takes a special relevance in games being, as said, software products which production is extremely costly. Early evaluation of prototypes allows earlier evaluation and consolidation of design decisions, which is an indispensable requirement for games.

But HE of games does not deal only with aspects traditionally covered by the human-computer interaction studies but with the fuzzy boundaries of the usability definition. Since the first uses of heuristics for evaluating games made by Malone [40] and Clanton [7], new heuristics deem to new aspects such as predicted fun and appeal. Clanton [7] extracted from the inspection of several games a list of heuristics divided into three modules: game interface, game mechanics and game playability. Game interface is identifiable by traditional human-computer interaction concepts but game mechanics and playability pointed to new aspects as narratives, characters, immersion, and engagement. Indeed, Clanton's work analysed how and why games engage users and paved the way for the application of heuristic evaluation for aspects such as fun and appeal, more related to the modern definitions of usability or user experience. Many authors consider that heuristics can be used to evaluate the overall user experience of games and propose methods to measure it [12, 14, 19, 31, 41, 55] as physiological measurements; expert evaluations; subjective, self-reported measures; and usability tests [28].

However, despite its ability to adapt to different semantic questions of the usability definition, traditional heuristics are poorly apt to games: either they contravene in achieving a good game experience [25] or they do not cover important aspects of the games [14]. Nielsen declared that there should be domain-specific heuristics for specific products [45]. According to this, different proposals adapted existing heuristics or compiled new lists to cover new aspects of games such as mobility [31] or multiplayer interactions [32]. Among these adaptations, playability measurement is one of the most challenging areas. There are various proposed heuristic guidelines focused on playability evaluation, but these proposals differ quite substantially from each other since these attempts are just emerging [30]. Comparisons between different playability heuristics sets [30, 33] when applied to evaluations are expanding the knowledge in this area.

Nonetheless and as stated before, few of the new proposals are used systematically. Nielsens' heuristics remains the most popular guideline [58] despite not being specially suitable for games and even considering the fact that computer games contravene Nielsen's guidelines and the ways these contraventions impact on the flow [25]. From a more conservative point of view, it can be considered that the majority of Nielsen's heuristics can be helpful when analyzing the interface of a game, but they mainly fails in the ability to address gameplay issues [14].

New proposed heuristic guidelines not only arise from the different usability aspects that need to be evaluated or the special characteristics of games, but also from the emergence of new platforms, genres or devices which allow new genres and interaction models that challenge usability evaluation. New heuristics guidelines that cope with these innovations have come to compile the best practices for future designs of every subset of games. These best practices are latterly used to assess usability on new prototypes belonging to the referred brand new type of games.

3 Proposal

Despite the existence of numerous new HE evaluation tools proposed in literature that cope with the different aspects of usability, the special characteristics of games and their innovative nature given the incessant emergence of different platforms, genres, devices, etc. make the academic community, as said before, still mainly rely on the Nielsen's heuristics as the preferred guideline when performing HE on games [58].

Hypothetically, the vast amount of different proposals specifically focused on every narrow subset of games and the explosion of these new subsets make it difficult for developers and evaluators to manage all the possibilities at the same time it hinders the selection better adapted to each case. Going further, the fragmentation and dispersion of the new proposals lead to the fact that, given a game and a list of heuristics, they can perfectly evaluate an usability aspect or particularities of the game genre while ignoring others that may potentially be covered by other heuristic list. Under these hypotheses, we propose that, given a type of game and the aspects of interaction to be assessed, we can offer a recompilation of the most convenient guidelines from the different existing proposals and rebuild them into a new one, a meta-heuristic, where non-applicable heuristics are minimized and the coverage of usability aspects and games specificities are maximized.

Given that the different proposals add different advantages, the added value seems obvious when combining different usability guidelines to obtain better results. But there is a risk when

the management of guidelines with a high number of items becomes impracticable. Additionally, the amount of useless items in terms of evaluation being rated "non applicable" will be enormous. The present proposal tries to establish mechanisms to customize the combination of guidelines to maximize covering of the usability aspects to evaluate and the particularities of the game under evaluation while minimizing non-applicable items.

The proposal of a meta-heuristic is based on the modular conception of heuristics already validated by different authors [11]. While Korhonen and Koivisto [31] proposed a modular guideline suggesting that every module could be used separately, allowing evaluators to focus solely on certain aspects of design, we extend this conception to potentially every heuristic guideline proposed in literature. These heuristic lists and their items can be considered as modules or pieces, which can be rearranged depending on the circumstances of the evaluation. Heuristic items are in many cases compilations of best practices extracted from literature, games reviews or experts' opinions under specific criteria determined by the researchers. In our proposal these criteria cross several proposed heuristic guidelines to select the applicable items. These selection criteria are related to the characteristics of the game under evaluation, to the usability aspects we want to evaluate and other circumstances of the evaluation.

Thereby, the development phase should also be explicitly considered when deciding if a heuristic item is selectable. The existence of playable prototypes does not impose important constraints to the evaluations. While "*gameplay should be evaluated already in an early design phase when there are design documents available*" [31], other aspects such as control design and interface feedback are only evaluable through a more sophisticated prototype.

To perform the rebuilding of heuristics and the application of selection criteria we propose MUSE (Meta-heUristics uSability Evaluation tool), a tool capable to help game developers and academic researchers to obtain better-adjusted lists for the evaluation of games.

Additionally, another reason to justify the discarding of specific heuristics in favor of the most popular generic guidelines is the lack of assistance from usability experts when performing evaluations. Frequently, developers or stakeholders with no deep knowledge in usability are the ones performing usability evaluations because of availability and cost restrictions. This reuse of human resources is convenient in many cases and may be enough to perform preliminary analysis or even to get good final results as it is estimated that even non-expert evaluators can achieve very good results [57]. Thereby, MUSE should facilitate the comprehension of the heuristic items to non-experts users who do not need to know the original sources from where the heuristics were compiled. The questions should be self-contained and posed in a granularity level of abstraction such as to allow non-expert evaluators answering them while only requiring basic knowledge on the subject.

According to this, the formalization of final goals of MUSE can be detailed into the next objectives:

1. The tool should provide a heuristic guideline customized for a given usability evaluation plan which includes the evaluation goals, the development status of the product, and the characteristics of the target game to be evaluated.
2. The tool should present the guidelines in a level of abstraction low enough to be useful to non-expert evaluators.
3. To offer advantages, the obtained guideline must offer better results when evaluating games than those of the most used heuristic list, Nielsen's, in terms of discovering of usability issues and in terms of validity, reliability and easiness of use.

4. The tool should produce coherent results minimizing false positives (items inappropriate to the context) and false negatives (useful items non selected into the customized compilation offered).

4 Design and development of MUSE

In order to achieve the goals enumerated in the previous section it is necessary to choose a subset of heuristic guidelines from the literature to work with. The next step is to add metadata tags to the different items from those heuristics in order to define a taxonomy that allows the selection of elements that meet given criteria. As stated before, these criteria are defined by the characteristics of a particular usability evaluation plan: which usability aspects will be evaluated, which are the characteristics of the specific game (genre, platform, etc.) and which is its current development phase.

The taxonomy derived from metadata tags will be applied only to the subset of heuristics selected from the literature but should be defined keeping in mind its potential applicability to any proposed heuristic guideline for games. This is important to facilitate the future evolution and growth of the tool.

It is necessary to extract similarities and differences from the selected heuristic guidelines, which will be part of the mentioned taxonomy. These similarities and differences crossed with the previously declared objectives will be the core of the proposed tool.

The process model chosen for the development of MUSE is based on a spiral model. It is important to remark that the goal of this paper is to present a first prototype that makes possible the future evolution of the tool. This first iteration has a limited scope so that efforts are focused on consolidating basic design decisions that will be fundamental in the future. The first proposed prototype will only partially cover the objectives formalized in the previous section but should take them into account explicitly and exhaustively. For instance, although the fourth goal previously defined is more ambitious and includes the minimization of false positives and false negatives, this preliminary version of the tool will maximize false positives with the inclusion of non-applicable questions as a trade-off to avoid loss of potentially interesting questions (maximizing false positives as a consequence of minimizing false negatives). In next iterations on the spiral lifecycle of the tool this aspect should be redefined and improved.

4.1 Selection of heuristics

The selection of heuristics should include some of the most referred heuristics adapted to games such as the works of Federoff [14], Korhonen and Koivisto [31], Pinelle, Wong and Stach [48] and Desurvire et al. [12]. They were chosen to start a summarization able to highlight similarities and common characteristics that will provide potential metadata in the subsequent taxonomy. Additionally, other works by the same authors were added when they are evolutions of first proposals either to cover new characteristics or to overcome detected deficiencies [11, 32, 49].

Additionally to these popular heuristics for games and in order to include works with a variety of scopes and objectives, some additional heuristics guidelines were added to the selected subset. These works are specifically focused on new genres or platforms [28, 47]. This way the final selection encompasses a representative subset of games.

It is important to remark again that this is an initial selection of heuristics from which a first version of the taxonomy should arise, which nonetheless is conceived as expandable by either adding other heuristics or elements to the taxonomy.

Finally, the selected heuristics include 237 items covering different aspects of usability, different goals and some distinct types of games as will be described in what follows (see Tables 1 and 2 for a summary):

- Federoff, 2002. One the most famous approaches to how heuristic evaluation can be undertaken over games is the work made by Federoff [14]. This compilation was extracted from literature, derived from the experience of the author in a day-to-day analysis in the industry and from the author's own knowledge. The goal of Federoff's compilation is to evaluate fun in any game genre. The heuristics were presented classified according to Malone's [40] categories and Nielsen's heuristics.

- Korhonen and Koivisto, 2006 and 2007. Korhonen and Koivisto[31] proposed a modular heuristic which covers three aspects: game usability, mobility and game play. Their proposal offers independent use of every module according to the needs of the evaluation.

Game usability "*covers the game controls and interface through which the player interacts with the game*". Inside this module five of the heuristics are related to visual design, three to how navigation is arranged and how the character can be controlled, and the rest are related to other aspects like getting feedback or how the game guides the player. Mobility module applies only when the game runs on a mobile platform. Gameplay module "*deals with issues that arises when the player interacts with the game mechanics and story*".

In a later work [32], Korhonen and Koivisto add an extra Multi-player module focused on online games that are played from mobile phones. However, they add a disclaimer: "*even if these heuristics have been designed for games that are played with mobile phones, most of the issues that have been mentioned hold true in non-mobile games as well*".

They added a preliminary evaluation of their proposal through experimental heuristic evaluations developed under restricted conditions.

- Pinelle, Wong, and Stach, 2008 and 2009. Pinelle, Wong and Stach published two heuristic proposals for games. The first one in 2008 [48] comes from a vast analysis of usability problems reported during reviews of commercial video games. It considered general usability problems found in PC games, but did not address problems found in specific game types. A subsequent compilation published in 2009 [49] pays special attention to genre types. Indeed, authors declare that despite the huge variety of game interfaces, layout or interaction methods, games belonging to the same genre have many user interface similarities. This last study inspects 382 public reviews of networked multiplayer games covering six genres: strategy, shooter, Role-Playing Game (RPG), sports, simulation and a miscellanea category.

- Desurvire, Caplan, and Toth, 2004 and 2009. Heuristic Evaluation for Playability (HEP)[12] is a set of heuristics published in 2004 and focused on playability evaluation. The compilation comes from the literature and from experts' advice and is specifically tailored to evaluate video, computer, and board games.

Table 1 Summary of selected heuristic guidelines (I)

	Focus on	Genre	Number of items	Classification
Federoff, 2002 [14]	Fun	Any*	38	Clanton's and Nielsen's
Korhonen & Koivisto, 2006 [31]	Mobile games	Any	29	Gameplay, Game Usability, Mobility
Korhonen & Koivisto, 2007 [33]	Multiplayer interactions	Multiplayer	8	Multiplayer
Pinelle, Wong & Stach, 2008 [48]		Any	10	None
Pinelle et al..., 2009 [1]	Networked multiplayer games	Networked multiplayer games: Strategy, Shooter, RPG, Sports, Simulation, Action and Other	10	None
Desurvire et al, 2004 HEP [12]	Video, computer, and board games		43	Game play, Game story, Game mechanics, Game usability
Desurvire & Wiberg, 2009 PLAY [49]	Early phases of game development	Real-Time Strategy (RTS), Action Adventure and First-Person Shooters (FPS).	50	Game Play, Skill Development, Tutorial, Strategy & Challenge, Game/Story Immersion, Coolness, Usability/Game Mechanics and Controller/Keyboard
Papaloukas et al, 2009 [47]	Usability game play, game interface and one includes fun, curiosity and challenge.	New genre game like Wii Sports or Pet Society in the popular social networking Facebook	10	None
Koeffel et al, 2010 [28]			39	Game play/game story, virtual interface and tabletop specific

*Koeffel et al [28] consider that Federoff focuses on role-playing games more tan other genres

Table 2 Summary of selected heuristic guidelines (II)

	How was compiled	Platform	Evaluation/Verification
Federoff, 2002 [14]	Literature and industry	Any	No
Korhonen & Koivisto, 2006 [31]	Nielsen's review focused on mobile gaming + evaluation of a game	Some mobile-specific heuristics	Different number of experts evaluated five games (alpha version)
Korhonen & Koivisto, 2007 [33]	Evaluation of 3 commercial mobile multiplayer games	Mobile	Informal and brief playability evaluation of commercial multi-player games.
Pinelle, Wong & Stach, 2008 [48]	Analysis of game reviews from a popular gaming website (108 different games and included 18 from each of 6 major game genres.)	PC-games	Preliminary evaluation of the heuristics: five people used them to evaluate a demo version of a PC game. Each evaluator completed an open-ended questionnaire at the end of the study.
Pinelle et al, 2009 [1]	382 public reviews of networked PC games	PC-games	Ten evaluators used the heuristic versus Baker et al's groupware usability heuristics to evaluate two networked multiplayer games.
Desurvire et al, 2004 HEP [12]	Literature and review by several playability experts and game designers.		Evaluation of a Flash prototype of a game using the heuristic performed by 1 evaluator vs. Playtesting with 4 users
Desurvire & Wiberg, 2009 PLAY [49]	Based on the existing HEP, and modified based on discussions with developers from Activision, THQ, Relic, Pandemic, Avalanche, Disney, and Microsoft Game Studios.		The authors declare that "Several design teams have used PLAY heuristics over the past 2 years"
Papaloukas et al, 2009 [47]	Literature + observation of playtesting recordings		Three experts observed five playtesting recording and classify problems observed attending to the 10 heuristics
Koeffel et al, 2010 [28]	Literature	Tabletop (some of the heuristics)	Heuristic evaluations on several games were conducted and the resulting data compared to user experience-based game reviews.

The 43 heuristics of this proposal are divided into four categories: Game Play, Game Story, Game Mechanics and Game usability. Game play stands for "*the set of problems and challenges a user must face to win a game*". *Game Story covers* "*all plot and character development*". Game Mechanics involves "*the programming that provides the structure by which units interact with the environment*". And finally, Game Usability "*addresses the interface and encompasses the elements the user utilizes to interact with the game (e.g. mouse, keyboard, controller, game shell, heads-up display)*".

In 2009 the Game Playability Principles (PLAY)[11] were published, a broad list of heuristics based on HEP and modified after discussions with developers from the industry. The authors declare it is a "*generalized foundation that could then be modified for each specific game*". The principles are grouped into seven categories: Game Play, Skill Development, Tutorial, Strategy & Challenge, Game/Story Immersion, Coolness, Usability/Game Mechanics, and Controller/Keyboard.

PLAY covers three genres: Real-Time Strategy (RTS), Action Adventure and First-Person Shooters (FPS).

- Papaloukas, Patriarcheas, and Xenos, 2009. The continuous emergence of new genres and platforms challenges the existing methods of usability evaluation. Papaloukas, Patriarcheas and Xenos[47] proposed a heuristic guideline specifically designed to address the new interaction models arisen from games inserted into social network sites and from games that use specific peripherals, such as the Wii console.

The authors defined usability as "*the degree to which a player is able to learn, control, understand, be intrigued and enjoy a game*" but the final compilation includes items mainly related to gameplay usability, game interface and only one applies to fun considerations. It is interesting to remark their conviction that, into the scenario of continuous novelties, "*heuristics can be developed for specific videogame categories by evaluating existing titles of video games, and by developing principles that describe the usability problems that might occurred*". Thus, their proposal is reactive to the emergence of new games but contemplates the fact that new products will emerge.

The proposed final compilation is extracted from the literature and also from observation of playtesting recordings.

- Koeffel et al., 2010. The Koeffel et al. [28] proposal offers 39 heuristics divided into three sets: game play/game story, virtual interface and some tabletop games specific heuristics. Several heuristic evaluations were conducted using the proposed framework to estimate the usefulness of the proposal, comparing the resulting data to experiences reported within the game reviews.

4.2 Filtering of the heuristic guidelines

The previous selection of heuristics guidelines gives rise to a set of 237 heuristics items. In order to provide a filtering method to pick out the most appropriate heuristic items from the whole set of 237 heuristics according to particular criteria, it is necessary to apply some metadata tags to each item. Some data preprocessing is required before the addition of metadata tags:

- Duplicated items or slightly different items obviously referred to the same question were removed.
- When items are composed by several questions in their original presentation and these questions can be answered independently, they were split into new items; if the presence of several questions is merely an instrument to provide a better explanation, no changes were made.
- Finally, the wording was homogenized to provide a coherent final set of heuristics.

After this preliminary filtering, a set of 160 heuristic items was ready to be tagged with metadata.

4.3 Definition of metadata

The definition of metadata tags should take into account that the items for a particular evaluation of a specific game will be chosen based on the underlying tagging. Thus, tags should characterize the potential evaluations and games while keeping correspondence with the available heuristic items, the information managed by these items and the focus of the original compilation.

The final taxonomy built from the metadata tag cloud should deal with the three main questions from the first goal formalized before:

(a) Which aspects of usability are subject of evaluation and which are the goals of the usability analysis,
(b) Which type of game will be analyzed attending to the subset of characteristics which determine a different focus of evaluation and,
(c) Which is the development status of the game regarding its degree of completion.

These three points will lead to different tag sets to be applied to the heuristic items allowing the final filtering. However, as stated before, the definition of every subset of tags should keep a correspondence with the available information from the managed heuristic guidelines.

In order to allow users to define which aspects of usability they are interested in evaluating (a) the final criteria was to maintain an approach coherent with the ISO definition of usability [23] and based on the three aspects of efficiency, effectiveness and satisfaction. Additionally, the classification by Clanton [7] was selected to describe the evaluation goals due to its frequent use in the selected heuristic guidelines. Clanton divided usability issues into three areas: game interface, game mechanics, and game play, which are partially related with the three aspects from the usability definition. Game interface refers to the device through which the player interacts with the game, corresponding to the issues derived from the interface in its basic conceptualization -buttons, menus, controls…- and consequently can be related to the efficiency aspect from the ISO definition. Game mechanics are the physics of the game, which are developed through a combination of animation and programming. Game mechanics are more related to how to do things than actually doing them, so it can be related to the effectiveness aspect. Game play is the process by which a player accomplishes the goal of the game and it is the issue most related to satisfaction. But if the boundaries of the concepts are fuzzy, during the processing of metadata it was confirmed that keeping both tag subsets —those based on the ISO's definition and those based on Clanton's classification— is required, since each one provides different and necessary information even though in many

cases the correlation works perfectly. Further analysis of evaluations developed using MUSE would presumably offer interesting data about how to classify the different approaches and define usability on games more accurately.

Also answering the question a), Nielsen's heuristics were added as categories. In most proposals they were used as categories in the first presentation of the new guidelines and hypothetically they can be useful for expert users when using the goals from their evaluation plan as selection criteria.

Another important aspect when evaluating usability is accessibility. Although accessibility questions can be subsumed under the category *Game Interface*, universal inclusion requirement seemed to justify the inclusion of an explicit tag. Unfortunately, only one of the heuristics items managed is closely related with accessibility questions. Future work needs to explicitly cover this area.

Definition of metadata tags required describing games with attention to characteristics relevant for usability assessment (b) leads to several problems. One of the most common ways to classify games is using the concept of genre inherited from other fields such as cinema or literature. The question of how to define genres in videogames has been widely discussed and it is controversial. But while some approaches reused the methodology of categorizations traditionally applied to films based in iconography, structure and theme, many authors agree that games have a particular characteristic not shared with other artistic productions [1, 56]: their ergodicity. Under this focus, the main aspect that can group similar games is interactivity [42]. "*The game's objective is a motivational force for the player, and this, combined with the various forms of interactivity present in the game, are useful places to start in building a set of video game genres*"[56] This point of view prioritizes questions about interface and controls rather than theme, classification of the storyline into the literature genres, or any other potential taxonomy.

Additionally to the genre, aspects such as platform, mode and milieu [1, 27] have been debated as possible categories. Platform, mode and milieu categorize games according to the hardware system and devices used, the environmental and experiential factors related to the spatial and temporal arrangements of the game, and the visual genre of the game, respectively. From these aspects only platform and (partially) the notion of mode can add relevant information when the goal is usability assessment.

Returning to the interactivity-based conception, Mark J. P. Wolf [56] proposed 42 genres to classify games. However, many cross-listing options are declared since the boundaries are considered fuzzy. The first attempts of using these categories as metadata tags demonstrated that there is a wide gap between the granularity of the heuristic items and the definition of the genres, so if this approach was selected, the final users of the tool would face unnecessary problems when declaring the genre of the game to be evaluated. These problems were overcome using a fine-grained classification in an abstraction layer under genres, which abandons any attempt of general classification and specifically focuses on questions closely related to what heuristic items evaluate. These more specific tags not only allows the evaluation of interactivity aspects but also platform details, relevant when evaluating usability, or the inclusion of questions about interaction between players when the game is multiplayer. This last analysis can be understood as related to the concept of mode.

This change of point of view to a fine grained tags-based description facilitates the addition of metadata to the heuristics and, as stated before, provides better management, updating, and extension capabilities while being a user-friendly approach since it avoids the need to know genre categorizations and genre cross-listing exceptions. Additionally, this approach leaves

open the possibility for further future development of the tool where rating of evaluation could be weighted according to the most relevant aspects to the evaluators. For instance, some evaluators could be more interested in the evaluation of interruptions rather than customization capabilities for the game, or in the measurement of immersion rather that in how well goals of the game are defined.

Finally, regarding development status (question c) the most important question is what is the degree to which the deliverable prototypes approach the final product. Although the different development methodologies define in a very different manner the process of building a game and its phases, a brief and coarse-grained approach is to distinguish between design phases where low-fidelity prototypes are the outcomes from development phases where higher levels of fidelity of the prototypes approximate the final product. While the categorization in early and late development stages suffers from fuzzy boundaries, it is a convenient way to allow selection of questions applicable to low-fidelity prototypes to those applicable to medium or high fidelity prototypes. At this stage of development of MUSE, no further complexity is required.

The final cloud of tags is shown in Fig. 1. They are classified according to the three questions to be covered and divided into seven categories. To take into account the aspects of usability subject of evaluation (a), the categories *Game Aspects* and *Usability Aspects* define the usability tests and *Nielsen's heuristics* is used to classify correspondences between guideline's items and Nielsen's proposed heuristics. Regarding to the characteristics of the game evaluated (b), *Platform* characterizes the platform where the game is played, *Purpose* allows a preliminary way to prioritize aspects according the final goals of the game, and *Keywords* is a miscellanea category where to merge different useful tags. Additionally, *Development Phase* is the category used to describe the development status of the game (c).

This classification of tags facilitates management, future update and extension, and mainly it aims to provide a friendlier interaction and experience for the users of the tool. Consequently, although "*Purpose of the game*" is a group of tags of limited utility in this preliminary version of the tool, they are introduced to allow easy extensibility by adding heuristics from the literature focused on educational games, serious games, etc.

According to the described categorization of metadata, 50 tags are the result of the analysis. They were applied to the final compilation of heuristics obtained from the processing described in the previous section. The addition of metadata was performed by one of the researchers and checked by the other three authors.

4.3.1 Tool modeling

According to the analysis presented above, the map of tags, the compiled list of heuristics and their relationships were implemented as a database, which allows a basic exploitation of data and some grade of achievement of the four defined objectives declared with minimal effort.

In the final model, the relationship between heuristic items and the defined tags follows the pattern many-to-many: any item can be related to an arbitrary subset of tags, each of them belonging to any specific category (see Fig. 1). For instance, the heuristic item "The main computer interface in pc game is hiding during game play?" matches with the keyword "Interface", the platform "PC Desktop", its focus of evaluation according to Clanton's classification is "Game Interface", the usability aspect from the ISO definition that it evaluates is "Efficiency", it can be matched with the 5th and 8th Nielsen's heuristic, and it will likely be

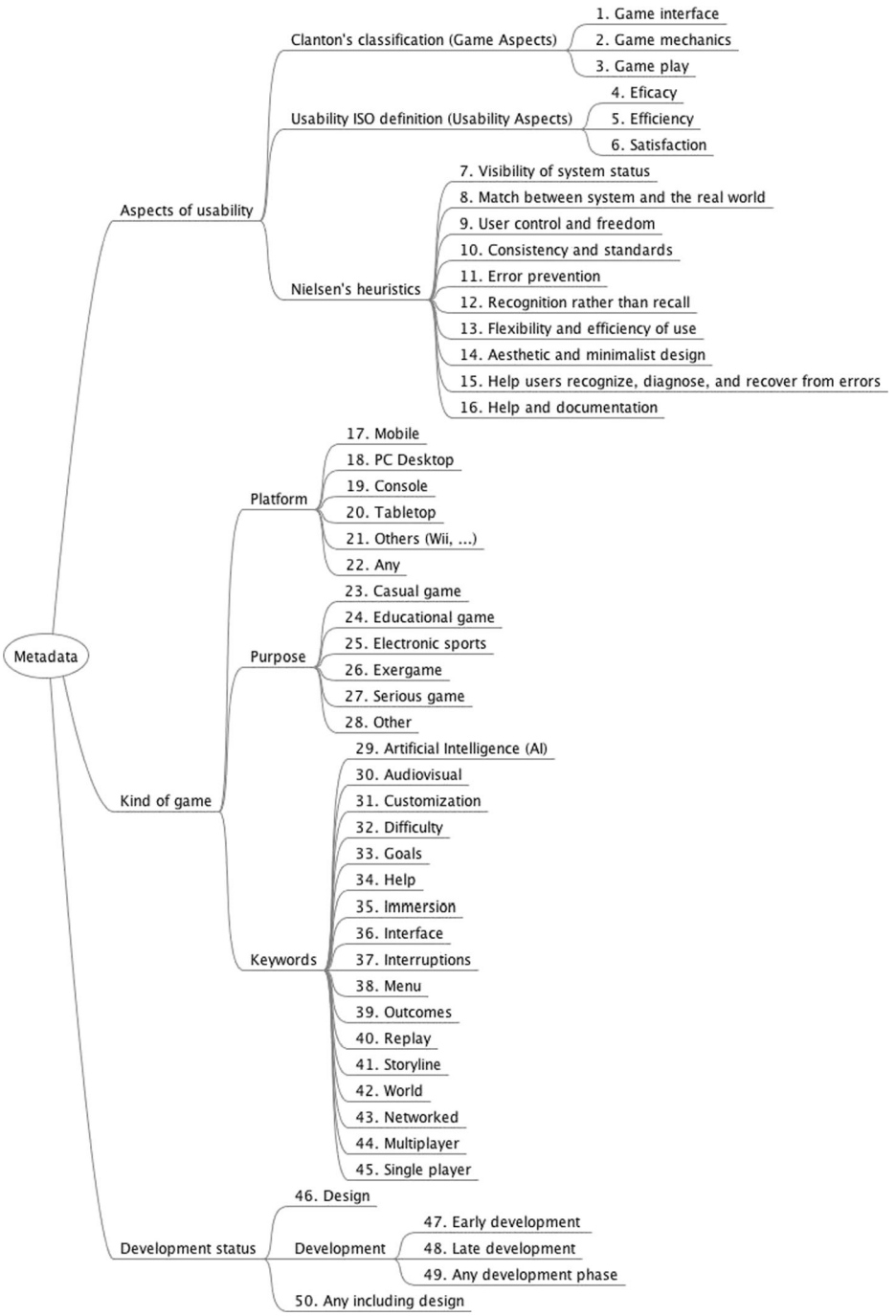

Fig. 1 Final cloud of metadata tags

more useful in "early development" phases although is applicable to "Any including design" development phase.

The described design leads to a preliminary prototype of the tool where probabilities of selecting heuristics items are maximized because of the matching with more than one tag. Although the fourth goal previously defined is more ambitious and includes the minimization of false positives and false negatives, in these preliminary version of the tool is accepted to increase the occurrence of non-applicable questions as a trade-off to avoid loss of potentially interesting questions (maximizing false positives as a consequence of minimizing false negatives). This seems to be a convenient approach since fine-grained adjustments can be done in later versions when data collected from evaluations will be available to drive a refinement of the selection criteria.

The database model is shown in Fig. 2. All the relationships between the heuristics and the map of tags follow the pattern many-to-many, so according to the third normal form, from the normalization rules established for databases, seven intermediate tables were created to interrelate items and all required constraints were defined to ensure referential integrity and to minimize data duplication.

This first prototype of MUSE will be merely used to identify potential gaps in the map of tags, the selected heuristic lists and their relationships. Although the language to query the tool will be SQL, which is closely related to the final logical model of the database, it provides a raw way to interact with the tool for accomplishing these testing objectives. Thereby, the parameters to customize the heuristic guideline, namely the tags selected, should be translated to these statements in SQL language. Because of this, the potential users of the first prototype of the MUSE tool are somehow restricted to its own developers, those who have the required insight of the date model to define proper SQL queries to interact with it.

5 Preliminary evaluations

Once a preliminary prototype of the tool is generated and facing the objectives presented in Section 2, there are many areas of evaluation pending of an exhaustive analysis. This first iteration should cover the cited objectives in some degree to later allow planning the evolution of the tool in future versions. The pending test batteries are out of the scope of this first approximation. A statistically relevant study of the validity of the tool and its characteristics

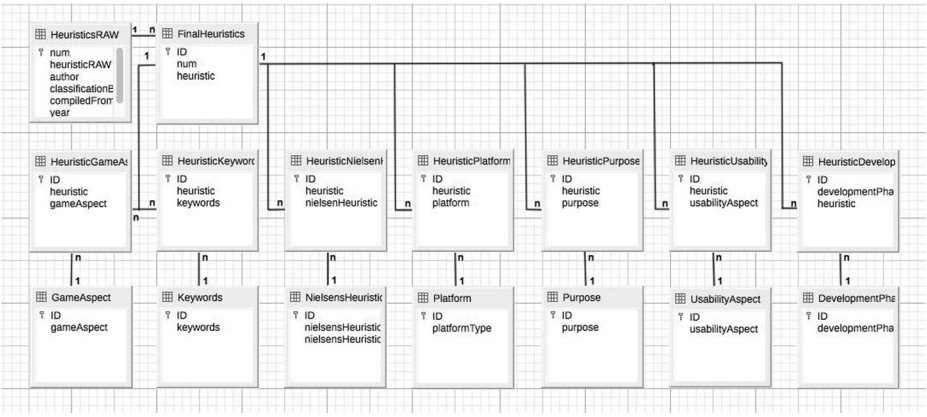

Fig. 2 MUSE database diagram

when compared to other tools needs a set with a high number of samples and tests designed according to the different objectives. Some sketches of the desirable future tests are described in the next section.

The first and main goal —*The tool should provide a customized heuristic guideline for a given usability evaluation plan that comprises the evaluation goals, the development status of the product, and the characteristics of the target game to be evaluated*— was the main driver for the design, as is described in the discussion exposed in Section 3, so it is mostly covered. Future work can update and extend the heuristics used as base for recompilation, as well as the map of tags, even the categorization tree. Indeed, the capability of extension was considered during the modeling of the tool and is part of the natural spiral-like evolution of the present proposal.

With regard to the second and third goals—*The tool should present the guidelines in a level of abstraction low enough to be useful to non-expert evaluators* and *To offer advantages, the obtained guideline must offer better results when evaluating games than those of the most used heuristic list, Nielsen's, in terms of discovering of usability issues and in terms of validity, reliability and easiness of use*—they are subject of the preliminary evaluations described in this section.

The fourth goal —*The tool should produce coherent results with no false positives (items inappropriate to the context) and false negatives (useful items non selected into the customized compilation offered)*— is left for future work because of the need of a higher number of evaluations and results to establish the level of accomplishment and it will potentially lead to fixes and adjustments. However, these evaluations already suggest that this is one of the first improvements that must be undertaken.

5.1 Pilot test

The objectives of the enclosed preliminary pilot test are checking the functionality of MUSE when rebuilding heuristics and testing how the MUSE's performance measurement can be done. This pilot test is based in the usability evaluation of two games: Grim Fandango Remastered [17] which can be classified as an Adventure game according to the classification proposed by Wolf [56] and Shadowmatic [53], a mobile game which can be classified as Puzzle under the same classification. These two games have been selected based on the different platforms they run on: Grim Fandango is designed for PCs while Shadowmatic is devised for mobile devices. Additionally, the first one relies on an avatar and storyline while Shadowmatic is a puzzle game with no story or avatar. These differences generate two distinct lists after applying every specific subset of tags to perform the recompilation of heuristics. However, an exhaustive analysis must require the evaluation of a higher number of games that cover almost every possible subset of tags from the defined taxonomy.

The two selected games were evaluated using both the Nielsen's heuristics and a purposely generated customized heuristic lists using MUSE in order to get an approximate measurement of the improvement degree at detecting usability issues. A volunteer member from the authors' institution performed the evaluations. The evaluator —male, 34 years old, gamer "every month"— is an IT professional but only has very basic notions of usability. This lack of expertise provides a preliminary overview of the level of accomplishment of the second objective.

The evaluator first applied Nielsen's heuristic and then the proposed heuristic to both games. This order may produce a bias, which should be considered in the analysis of the results.

The outcomes for the four heuristic evaluations in terms of issues detected and their severity rating according to the evaluator are shown in Figs. 3 and 4. From the analysis of Grim

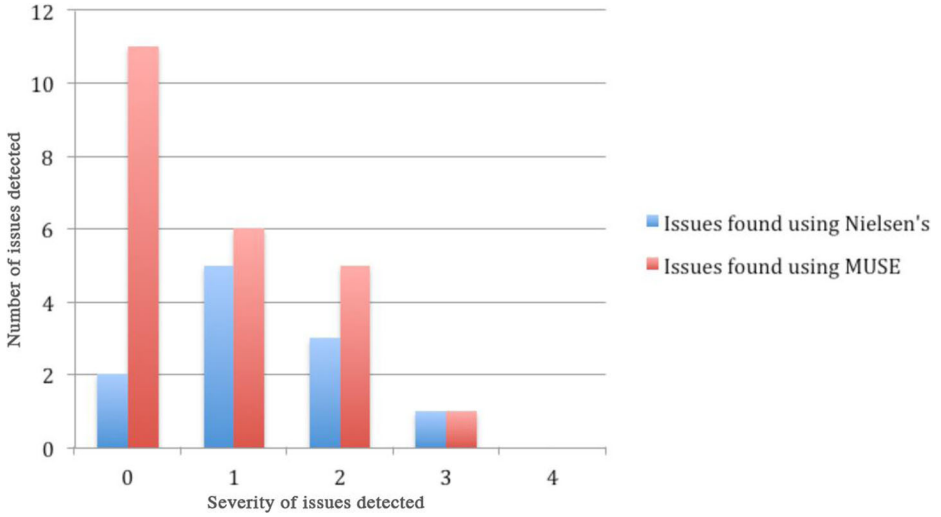

Fig. 3 Grim Fandango's evaluation using Nielsen's heuristic versus MUSE

Fandango, 11 usability issues were detected using Nielsen's heuristic whereas 23 were detected using MUSE. From the evaluation of Shadowmatic 9 usability issues were found using Nielsen's heuristic versus 25 issues using the proposed tool. There were no relevant differences in severity of issues detected (x-axis on figures) so we can hypothesize that the tests have a similar depth of analysis despite of the differences in the granularity of the questions. It is interesting to note that a 33% and a 17% respectively of the usability issues detected using the heuristics from the customized compilation were already discovered using Nielsen's heuristics. The rest of items, 67 and 83% of the total respectively, are new and different from the issues detected in the first analysis.

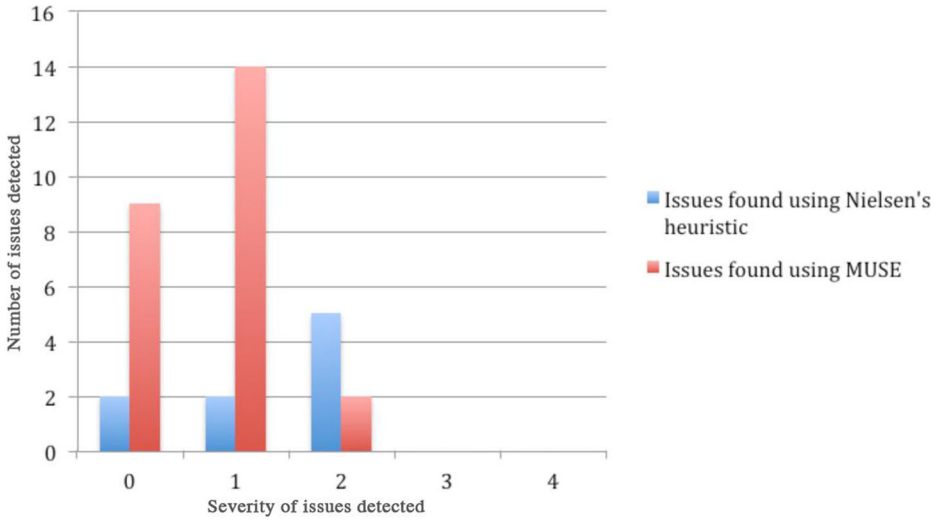

Fig. 4 Shadowmatic's evaluation using Nielsen's heuristic versus MUSE

It is also important to note that, although an experienced evaluator will likely use Nielsen to ultimately consider similar questions that those proposed by MUSE, a non-expert evaluator with less experience and knowledge will use both tools as independent lists. Nonetheless, Nielsen's heuristics are not explicitly included into the heuristics rebuilt by MUSE as they are only used as categories, as stated before, and it can be considered that according to the level of abstraction of questions both tools are independent. A deeper analysis capable to demonstrate statistically the independence of the two heuristics and potentially the independence of the rebuilt meta-heuristic versus the heuristics used as core in the generation is out of the scope of this paper.

Additionally to the usability evaluation itself, an interview with the evaluator offered information about feelings regarding easiness of use and usefulness of the different heuristic lists. Being the evaluator a non-expert in usability, he found it difficult to apprehend the full scope of Nielsen's heuristics, which are stated in a more abstract redaction. The evaluator declared that the lists from MUSE were easier to apply because of their granularity.

These preliminary results suggest that second and third goals can be covered to some extent but obviously it is necessary to run formal evaluations.

A future improvement is related to the fourth goal. Whereas the existence of false negatives (useful items non included in the customized compilation) was not analyzed, the existence of false positives is easily measurable using as metric the number of items evaluated with a "Not applicable (NA)" answer. There are a 21% of items marked with NA in the case of Grim Fandango and a 22% for Shadowmatic. We consider there is potential for improvement in this ratio and such improvement should be addressed in next versions of the tool.

Besides demonstrating that the prototype is functional, this pilot test also leads to an interesting question: the evaluation of second and third goals is a challenging area and new metrics needs to be introduced in future evaluations.

5.2 Preliminary usability evaluation

In order to deepen in the evaluation of objectives, and according to the results of the pilot test, new evaluation metrics of evaluation need to be considered. To introduce a complementary quantitative measurement of achievement of the goals, the "evaluator effect" has been introduced into the evaluation plan.

The evaluator effect refers to the fact that multiple evaluators evaluating the same interface with the same evaluation tool detect different sets of problems. This effect exists "for both novice and experienced evaluators, for both cosmetic and severe problems, for both problem detection and severity assessment, and for evaluations of both simple and complex systems"[20]. The evaluator effect should be minimized to improve the reliability of evaluations, distinguishing reliability, the extent to which independent evaluations produce the same results, from validity, the extent to which the problems detected during an evaluation are also those that show up during real-world use of the system.

The evaluator effect can be minimized by avoiding vague goal analyses, vague evaluation procedures and vague problem criteria. However, the best practices proposed as systematization of evaluation such as sampling the scenarios to be evaluated, the exhaustive description of the tasks or the definition of specific problem criteria are part of the challenge of evaluating games as stated in the Introduction. All these best practices are difficultly applicable to games. However, MUSE was built using fine-grained questions, which establish by default some

Table 3 SUS-like questionnaire applied to the evaluation tools

	SUS-like questions
1	I think that I would like to use this tool frequently
2	I found the tool unnecessarily complex.
3	I thought the tool was easy to use.
4	I think that I would need the support of a technical person to be able to use this tool.
5	I found the various functions in this tool were well integrated.
6	I thought there was too much inconsistency in this tool.
7	I would imagine that most people would learn to use this tool very quickly.
8	I found the tool very cumbersome to use.
9	I felt very confident using the tool.
10	I needed to learn a lot of things before I could get going with this tool.

restrictions for the evaluators. According to this it can be hypothesized that concretion of the items should lead to a lower evaluator effect when comparing with the use of Nielsen's heuristic list in this specific application area.

Additionally, the evaluation of the third goal focused on the easiness of use of the tools, namely the usability of the usability evaluation tool. To measure this aspect a rewrite of the popular test SUS [3] was used as a questionnaire well addressed to obtain the evaluators feedback (see Table 3).

The design of the test followed an Inner Group [36] approach: half of the evaluators run Nielsen's heuristic first and the other half used MUSE generated heuristic first. Evaluators E1, E2, E3 and E5 belongs to the first group who used Nielsen's first; E4, E6, E7 and E8 belongs to the second group. No significant differences found when comparing the two groups.

This test was performed over two games: Grim Fandango Remastered [17], which licenses were freely provided by the developers of the game, and Roll the Ball [51], which was selected instead of Shadowmatic because being available free of charge. Again these two games were selected because they both differ in their target platform being Grim Fandango a PC game and Roll the Ball one for mobile devices, as well as Grim Fandango is avatar-based and follows a storyline while Roll the Ball is a puzzle game with no story nor avatar.

Eight evaluators performed the evaluation following the recommendations of Hwang and Salvendy [21] and increasing in two the number recommended by Nielsen [45]. However, it is necessary to insist on the fact that to perform statistically relevant tests we need a bigger sample set. The eight evaluators participated voluntarily obtaining as a gift the Grim Fandango's license required to download and activate the game. The evaluators were recruited between IT professionals and none of them had previous knowledge on usability beyond the basics so the group is homogeneous regarding their level of expertise, which is especially relevant in order to compare the evaluator effect. The evaluators are between 34 and 41 years old, all of them are men and regarding their familiarity with games three of them declared to play games "every day", three play "every week" and two "every month". The fact that all evaluators are gamers has probably occurred because in the recruiting phase the volunteers were informed that the evaluation would be run over games, and because the compensation is in form of a game license.

The evaluator effect was measured following the metric *any-two agreement* proposed by Hertzum and Jacobsen [20] which measures to what extent pairs of evaluators agree on what problems the system contains:

Table 4 Any-two agreement between evaluators on Grim Fandango's evaluation using Nielsen's heuristic

GrimFandango/Nielsen	E1	E2	E3	E4	E5	E6	E7	E8
E1	x	0	0	0	25	0	0	7.69230769
E2	x	x	9.52380952	10	0	0	0	11.1111111
E3	x	x	x	6.66666667	5.55555556	0	0	0
E4	x	x	x	x	14.2857143	0	0	0
E5	x	x	x	x	x	0	11.1111111	0
E6	x	x	x	x	x	x	7.69230769	5.26315789
E7	x	x	x	x	x	x	x	0
E8	x	x	x	x	x	x	x	x

$$Any{-}two\ agreement = Average\ of\ \frac{Pi \cap Pj}{Pi \cup Pj}\ over\ all\ {}^{1}/_{2}n(n{-}1)\ pairs\ of\ evaluators$$

Where P_i and P_j are the sets of problems detected by evaluator i and evaluator j, respectively, and n is the number of evaluators. The calculated any-two agreement values in the test for each game are shown in the Tables 4, 5, 6 and 7. A summary of results is shown in Table 8. According to Hertzum and Jacobsen [20], the average agreement between any pair of evaluators who have evaluated the same system using the same evaluation tool ranges from 5 to 65%. According to the results shown in Table 8, MUSE reduces the evaluator effect with accordance levels in the high extreme of the range while Nielsen's heuristic, in the circumstances of the experiment, leads to accordance levels on the low extreme of the range.

With regard to false positives, the number of items inappropriate for the context marked as Non-applicable by the experts when using the rebuilt heuristic by MUSE, averaged of 11,8% (standard deviation of 5) for Grim Fandango's evaluation and an average of 13,4% ($\sigma = 4{,}7$) in the evaluation of Roll the Ball. These high values added to a 3,7% ($\sigma = 1{,}8$) of non-answered or confusing items in the case of Grim Fandango and 2,6%($\sigma = 1{,}8$) for Roll the Ball, giving an extremely high number of non-significant items for the specific evaluation that may negatively impact on usability of the tool in terms of efficiency and satisfaction. The pilot test pointed to the need to improve this aspect but, after the evaluation, it is confirmed that should be marked as a priority improvement in next iterations of the tool.

Regarding the usability of the tools themselves, every evaluator was provided with two questionnaires based on a rewrite of the SUS test, each one corresponding to each tool (see

Table 5 Any-two agreement between evaluators on Roll the Ball's evaluation using Nielsen's heuristic

Roll the Ball/Nielsen	E1	E2	E3	E4	E5	E6	E7	E8
E1	x	6.25	5.55555556	0	0	8.33333333	12.5	8.33333333
E2	x	x	4.54545455	12.5	7.69230769	6.25	8.33333333	6.25
E3	x	x	x	16.6666667	13.3333333	5.55555556	7.14285714	5.55555556
E4	x	x	x	x	11.1111111	0	12.5	8.33333333
E5	x	x	x	x	x	0	0	0
E6	x	x	x	x	x	x	12.5	0
E7	x	x	x	x	x	x	x	0
E8	x	x	x	x	x	x	x	x

Table 6 Any-two agreement between evaluators on Grim Fandango's evaluation using MUSE

Grim Fandango/ MUSE	E1	E2	E1	E4	E5	E6	E7	E8
E1	x	0.61151079	0.58992806	0.58273381	0.55395683	0.61870504	0.5971223	0.6618705
E2	x	x	0.67625899	0.57553957	0.66906475	0.62589928	0.54748201	0.68345324
E3	x	x	x	0.63309353	0.58992806	0.64748201	0.71223022	0.64748201
E4	x	x	x	x	0.55395683	0.65467626	0.58273381	0.61151079
E5	x	x	x	x	x	0.56115108	0.57553957	0.5323741
E6	x	x	x	x	x	x	0.54028777	0.64028777
E7	x	x	x	x	x	x	x	0.66906475
E8	x	x	x	x	x	x	x	x

Table 3). Despite the scoring cannot be read in terms of percentiles as is usually made in the SUS post process, the final average of 57 for both tools ($\sigma = 18$ for Nielsen and $\sigma = 21$ for MUSE) leads to the preliminary conclusion that both tools can be considered similar in terms of effectiveness, efficiency and satisfaction of use.

Additionally, the evaluators were asked to complete the questionnaire with informal commentaries and suggestions. The majority of them expressed that the main difficulties they encountered using Nielsen were because the level of abstraction of the items and the problems distinguishing the scope of every heuristic. Regarding to MUSE, the evaluators declared that the number of questions posed was too big and that they encountered some difficulties when interpreting some of the heuristics. These results strength the importance of diminishing drastically the number of false positives in the rebuilding performed by the tool and, additionally, it highlights the usefulness of the addition of a brief explanation for every proposed question as help documentation for the evaluators.

Despite the similarity of results extracted from the SUS-like questionnaires, there was a significant difference in the number of usability issues detected when using one or another tool. As shown in Table 9, the evaluators detected significantly lower number of issues when they used Nielsen's. However, this difference probably could be explained by the bias introduced in the test by the lack of expertise in usability evaluation of the volunteers. The comparison of items detected using one or another tool is useless this time because the issues reported when using Nielsen's are declared in very generic and abstract sentences and the majority of them referring to general feelings about the interfaces. It is quite likely that a

Table 7 Any-two agreement between evaluators on Roll the Ball's evaluation using MUSE

Roll the Ball/MUSE	E1	E2	E3	E4	E5	E6	E7	E8
E1	x	0.61290323	0.58064516	0.580645161	0.61290323	0.67741935	0.58870968	0.62096774
E2	x	x	0.60483871	0.572580645	0.58870968	0.64516129	0.45967742	0.62903226
E3	x	x	x	0.491935484	0.55645161	0.54032258	0.41935484	0.58870968
E4	x	x	x	x	0.59677419	0.53225806	0.48387097	0.52419355
E5	x	x	x	x	x	0.60483871	0.55645161	0.62096774
E6	x	x	x	x	x	x	0.62903226	0.7016129
E7	x	x	x	x	x	x	x	0.54032258
E8	x	x	x	x	x	x	x	x

Table 8 Any-two agreement calculation summary

Any-two agreement	Nielsen		MUSE	
	Average	Standard Deviation	Average	Standard Deviatich
Grim Fardango	4.067	6.16	61.947	6.32
Roll the Ball	6.401	5	71.571	7.51

preliminary training for the evaluators on Nielsen's heuristic would have been very useful. Thus, this aspect of the evaluation does not lead to conclusive results.

This time evaluators were not asked to establish severity ratings for the identified usability issues so further analysis of the capability of MUSE on detecting severe usability issues is also subject of future work.

6 Conclusions and future work

In this paper, we presented MUSE, a tool for automatic creation of customized usability heuristic guidelines for evaluation on games. This tool is able to rebuild heuristics from literature to provide a customized set of heuristic questions taking into account the evaluation plan, namely the goals of the evaluation, the specificities of the game being evaluated and its development status. The objectives were formalized, design decisions were discussed and a preliminary evaluation was performed.

The results of the preliminary evaluation suggest that MUSE is a promising approach to improve the detection of usability issues when the evaluators are not experts. Additionally, the use of MUSE reduces the evaluator effect, increasing the reliability of the tool over other generalist tools when applied on games.

The evaluation results also suggest the relevance of a better control over the number of items inappropriate for the context, false positives, included in the rebuilt heuristic list obtained from MUSE. The design decision of minimizing false negatives even when this could favor the increase of false positives should be reconsidered. Additionally, it is also elicited from the experiment the need of an enclosed documentation, which helps evaluators to understand every question of the heuristic list.

Finally, the obtained outcome of this first iteration of the tool is a prototype of MUSE, which covers the defined basic goals. The cloud of tags resulting from the analysis of the literature and taken as core of the tool, was designed expandable and upgradeable and still could still be subject of further refinement. It is necessary to include new heuristics from

Table 9 Number of usability issues in every evaluation

Number of usability issues detected		E1	E2	E3	E4	E5	E6	E7	E8
Usign Nielsen	Grim Fandango	3	8	13	2	5	9	4	10
	Roll the Ball	6	10	12	6	3	6	2	6
Using MUSE	Grim Fandango	70	76	95	82	61	75	98	90
	Roll the Ball	68	51	46	42	66	78	85	68

literature and new categorizations to enrich it and improve its usefulness and versatility. Moreover, the future addition of new tags and metadata would also contribute to a reduction of false positives and false negatives, which is essential to reduce the checklist and made it efficient and manageable.

A bigger battery of evaluations is needed for every prototype of MUSE, while statistically significant evaluations could be very useful to consolidate and measure the achievement of the defined goals. From the first performed approximations it seems useful to combine qualitative and quantitative metrics into the evaluation of the tool itself.

Among the desirable expansions of the tool, taking accessibility into account is essential to achieve universal designs. This attention to inclusion is especially relevant with the popularization of serious games, which are by definition oriented to population with a wide range of sensory and cognitive particularities. Attending accessibility requires further exploration of the literature to add specific heuristic to the core of the tool or even to create new items to contemplate accessibility questions.

Additionally the spectrum of target games potentially evaluable is a set with flexible boundaries: gamification as a trend and the exploration of augmented reality, virtual reality and pervasive games add new variables to the concept of game. Customizations around these new gaming products would require the inclusion of specific heuristics into the core of the tool and reviewing the applicability of the already included heuristics. Some authors already discussed the use of heuristics by pervasive game developers [50], or the potential need for new guidelines specifically designed for smart home environments. The HEP [12] heuristics seem to be transferable to pervasive gaming applications in smart home environments.

Finally, rating is one of the issues to be improved in future versions. A classical yes/no/not applicable answering model was chosen but several authors proposed enriched methods for rating the results [37] that could be included in MUSE. Going further, evaluators can take advantage of the modular design of the tool by adding weights to the different aspects according to the evaluation plan, its priorities and goals. The metadata can also be enriched with these priorities and newly found derived issues can be offered in non-flat list form.

Acknowledgements This work has been partially supported by the EU project SmokeFreeBrain (PI055-15/E03) and by the Telefonica Chair "Intelligence in Networks" of the Universidad de Sevilla, Spain. The licences for the Grim Fandango Remastered copies used during the tests where kindly provided by the creator of the remastered version and owner of its Intellectual Property, Double Fine Productions.

References

1. Apperley TH (2006) Genre and game studies: Toward a critical approach to video game genres. Simul Gaming 37(1):6–23
2. Barendregt W, Bekker MM, Bouwhuis DG, Baauw E (2006) Identifying usability and fun problems in a computer game during first use and after some practice. Int J Human-Comput Stud 64(9):830–846
3. Brooke J (1996) SUS-A quick and dirty usability scale. Usab Eval Indust 189(194):4–7
4. Brown E, Cairns P (2004) A grounded investigation of game immersion. In CHI'04 extended abstracts on Hum Factors in Comput Syst (pp. 1297-1300). ACM

5. Caillois R (1961) Man, play, and games. University of Illinois Press
6. Cheng K, Cairns PA (2005) Behaviour, realism and immersion in games. In CHI'05 extended abstracts on Hum Factors in Comput Syst (pp. 1272-1275). ACM
7. Clanton, C. (1998). An interpreted demonstration of computer game design. In CHI 98 conference summary on Hum Factors in Comput Syst (pp. 1-2). ACM
8. Cowley B, Charles D, Black M, Hickey R (2008) Toward an understanding of flow in video games. Comput Entertain (CIE) 6(2):20
9. Csikszentmihalyi M (1975) Play and intrinsic rewards. Journal of humanistic psychology
10. Dempsey JV, Lucassen BA, Haynes LL, Casey MS (1997) An exploratory study of forty computer games. University of South Alabama, Mobile, AL
11. Desurvire H, Wiberg C (2009) Game usability heuristics (PLAY) for evaluating and designing better games: The next iteration. In International Conference on Online Communities and Social Computing (pp. 557-566). Springer, Berlin, Heidelberg
12. Desurvire H, Caplan M, Toth JA (2004). Using heuristics to evaluate the playability of games In CHI'04 extended abstracts on Hum Factors in Comput Syst (pp. 1509-1512). ACM
13. FDIs I. S. O (2009) 9241-210: 2009. Ergonomics of human system interaction-Part 210: Human-centered design for interactive systems (formerly known as 13407). International Organization for Standardization (ISO). Switzerland
14. Federoff MA (2002). Heuristics and usability guidelines for the creation and evaluation of fun in video games (Doctoral dissertation, Indiana University)
15. Frøkjær E, Hertzum M, Hornbæk K (2000) Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 345-352). ACM
16. Garzotto F (2007) Investigating the educational effectiveness of multiplayer online games for children. In Proceedings of the 6th international conference on Interaction design and children (pp. 29-36). ACM
17. Grim Fandango Remastered (2015) http://www.grimremastered.com/ Lucasfilm Ltd. Retrieved in November March 2017
18. Hassenzahl M (2008) User experience (UX): towards an experiential perspective on product quality. In Proceedings of the 20th Conference on l'Interaction Homme-Machine (pp. 11-15). ACM
19. Hazlett RL (2006) Measuring emotional valence during interactive experiences: boys at video game play. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 1023-1026). ACM
20. Hertzum M, Jacobsen NE (2001) The evaluator effect: A chilling fact about usability evaluation methods. Int J Human-Comput Interact 13(4):421–443
21. Hwang W, Salvendy G (2010) Number of people required for usability evaluation: the 10 ± 2 rule. Commun ACM 53(5):130–133
22. IJsselsteijn W, De Kort Y, Poels K, Jurgelionis A, Bellotti F (2007) Characterising and measuring user experiences in digital games. Int Conf Adv Comput Entertain Technol 2:27
23. Iso W (1998) 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs). The international organization for standardization, 45, 9
24. Järvinen A, Heliö S, Mäyrä F (2002) Communication and community in digital entertainment services. Prestudy Research Report
25. Johnson D, Wiles J (2003) Effective affective user interface design in games. Ergonomics 46(13-14):1332–1345
26. Ke F (2008) Alternative goal structures for computer game-based learning. Int J Comput-Support Collab Learn 3(4):429
27. Ki King G, Krzywinska T (2002) Screenplay: cinema/videogames/interfaces. Wallflower Press
28. Koeffel C, Hochleitner W, Leitner J, Haller M, Geven A, Tscheligi M (2010) Using heuristics to evaluate the overall user experience of video games and advanced interaction games. In Evaluating user experience in games (pp. 233-256). Springer, London
29. Korhonen H (2010) Comparison of playtesting and expert review methods in mobile game evaluation. In Proceedings of the 3rd International Conference on Fun and Games (pp. 18-27). ACM
30. Korhonen H (2011) The explanatory power of playability heuristics. In Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (p. 40). ACM
31. Korhonen H, Koivisto EM (2006) Playability heuristics for mobile games. In Proceedings of the 8th conference on Human-computer interaction with mobile devices and services (pp. 9-16). ACM
32. Korhonen H, Koivisto EM (2007) Playability heuristics for mobile multi-player games. In Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts (pp. 28-35). ACM
33. Korhonen H, Paavilainen J, Saarenpää H (2009) Expert review method in game evaluations: comparison of two playability heuristic sets. In Proceedings of the 13th international MindTrek conference: Everyday life in the ubiquitous era (pp. 74-81). ACM

34. Larsen JM (2008) Evaluating user experience–How game reviewers do it. In Evaluating User Experiences in Games, Workshop at CHI
35. Law E, Roto V, Vermeeren AP, Kort J, Hassenzahl M (2008) Towards a shared definition of user experience In CHI'08 extended abstracts on Hum Factors in Comput Syst (pp. 2395-2398). ACM
36. Lazar J, Feng JH, Hochheiser H (2017) Research methods in human-computer interaction. Morgan Kaufmann
37. Livingston IJ, Mandryk RL, Stanley KG (2010) Critic-proofing: how using critic reviews and game genres can refine heuristic evaluations. In Proceedings of the International Academic Conference on the Future of Game Design and Technology (pp. 48-55). ACM
38. Mack RL, Nielsen J (1994) Usability inspection methods. Wiley & Sons, New York, pp 1–414
39. Malone TW (1981) Toward a theory of intrinsically motivating instruction. Cogn Sci 5(4):333–369
40. Malone TW (1982). Heuristics for designing enjoyable user interfaces: Lessons from computer games. In Proceedings of the 1982 conference on Hum Factors in Comput Syst(pp. 63-68). ACM
41. Mandryk RL, Atkins MS (2007) A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int J Human-Comput Stud 65(4):329–347
42. Myers D (2003) The Nature of Computer Games: Play As Semiosis (Digital Formations;, V. 16,). Peter Lang Publishing
43. Nacke L (2009) From playability to a hierarchical game usability model. In Proceedings of the 2009 Conference on Future Play on@ GDC Canada (pp. 11-12). ACM
44. Nielsen J (1994) Usability inspection methods. In Conference companion on Human factors in computing systems(pp. 413-414). ACM
45. Nielsen J (1994) Usability engineering. Elsevier
46. Nielsen J, Molich R (1990) Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 249-256). ACM
47. Papaloukas S, Patriarcheas K, Xenos M (2009). Usability assessment heuristics in new genre videogames. In Informatics, 2009. PCI'09. 13th Panhellenic Conference on (pp. 202-206). IEEE
48. Pinelle, D., Wong, N., & Stach, T. (2008). Heuristic evaluation for games: usability principles for video game design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1453-1462). ACM
49. Pinelle D, Wong N, Stach T, Gutwin C (2009) Usability heuristics for networked multiplayer games. In Proceedings of the ACM 2009 international conference on Supporting group work (pp. 169-178). ACM
50. Röcker C, Haar M (2006) Exploring the usability of videogame heuristics for pervasive game development in smart home environments. In Proceedings of the third international workshop on pervasive gaming applications–pergames (pp. 199-206)
51. Roll the Ball, https://play.google.com/store/apps/details?id=com.bitmango.rolltheballunrollme&hl=es_419 Retrieved in March 2018
52. Sánchez JLG, Vela FLG, Simarro FM, Padilla-Zea N (2012) Playability: analysing user experience in video games. Behav Inform Technol 31(10):1033–1054
53. Shadowmatic (2015) https://www.shadowmatic.com/ Triada Studio Games Retrieved in November 2017
54. Susi T, Johannesson M, Backlund P (2007) Serious games: an overview. Technical Report HS IKI TR07001. School of Humanities and Informatics, University of Skövde, Sweden
55. Sweetser P, Wyeth P (2005) GameFlow: a model for evaluating player enjoyment in games. Comput Entertain (CIE) 3(3):3–3
56. Wolf MJ (2001) The medium of the video game. University of Texas Press
57. Yáñez Gómez R, Cascado Caballero D, Sevillano JL (2014) Heuristic evaluation on mobile interfaces: A new checklist. Sci World J
58. Yáñez-Gómez R, Cascado-Caballero D, Sevillano JL (2017) Academic methods for usability evaluation of serious games: a systematic review. Multimed Tools Appl 76(4):5755–5784