Field Playtesting with Experts' Constructive Interaction: An Evaluation Method for Mobile Games for Cultural Heritage

Panayiotis Koutsabasis
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0003-0478-7456

Anna Gardeli
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0001-9058-6516

Konstantinos Partheniadis
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0001-5421-7844

Panagiotis Vogiatzidakis
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0003-2898-0752

Vassiliki Nikolakopoulou
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0001-6410-6697

Pavlos Chatzigrigoriou
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0002-9574-0567

Spyros Vosinakis
Dept. Product & Systems Design Eng.
University of the Aegean
Syros, Greece
ORCID: 0000-0003-1735-4297

***Abstract*— Expert reviews of mobile games for cultural heritage (MG-CH) may predictively assess various aspects of game and content design in an iterative development process. They are not common because finding experts with knowledge for all aspects of MG-CH is difficult, while related heuristic sets are lengthy. Furthermore, ecologically valid evaluations of MG-CH must be performed in the field and examine various dimensions of playability, including gameplay, usability, functionality, and contextual factors about the location and the cultural content. To address those challenges, we have constructed a formative, empirical design review method of MG-CH, which requires experts to perform field playtesting in pairs, with minimal researcher intervention. We have refined the approach to collect data and insights about various aspects of playability; and then applied it in the evaluation of two MG-CH for UNESCO intangible CH sites that follow the same system and gameplay design. We found that the method yields many performance indicators and playability findings; facilitates participants to propose thoughtful recommendations; and affords in-situ production of cultural content.**

Keywords*— *evaluation method; mobile games; cultural heritage; playability; constructive interaction; co-discovery learning;

I. INTRODUCTION

Mobile games (MG) are pervasive games in which gameplay includes challenges that are based on the player's location (i.e. location-based mobile games). According to [3], in location-based MG some of the activity takes place in the physical domain and involves actions such as moving to a location, observing a physical object, taking pictures or recording sounds; at the same time, a part of the activity takes place in the digital domain where the players interact with digital content and may engage in problem-solving activities.

Cultural heritage (CH) is the legacy of a group or society that is inherited from past generations, and includes tangible culture (such as buildings, monuments, landscapes, books, works of art, and artefacts), intangible culture (such as folklore, traditions, language, and knowledge), and natural heritage (including culturally significant land-scapes, and biodiversity) (UNESCO).

Mobile games for cultural heritage (MG-CH) are a class of interactive mobile games that convey CH-related information and knowledge to visitors of CH sites, via various playful and engaging approaches and patterns based on the user location. MG-CH have several unique characteristics, including but not limited to:

- They are concerned with the creation of awareness and learning about the heritage of various cultural sites like GLAM (galleries, libraries, archives, museums), archaeological sites, cities or settlements of important heritage, monuments, UNESCO sites, etc.
- The game design and play are interweaved with multifarious CH content: narrative, characters, media, photographs, 3D models, challenges, hints, rewards, etc.
- CH content must be reviewed by heritage experts.
- The game must allow players (visitors of a CH site) observe sites at their own pace.
- People rarely visit cultural sites alone; players are anticipated to play in pairs or small groups.

The evaluation of these MG-CH often emphasizes usability and general user experience [22] that are important but not the sole dimensions of the player experience; other significant dimensions are related to mobile gameplay, the location and CH content. To some extent, these aspects can be revealed in early expert evaluation of MG-CH, before field testing with end-users, during an iterative design and development process. Finding single experts with knowledge and experience about all aspects of an MG-CH is difficult, if at all possible; thus a mixed group of CH professionals, game and interaction designers can be employed.

In this paper we construct and apply an approach of formative, empirical design review of location-based MG-CH which recruits experts to playtest early versions of the game following the principles of constructive interaction (or co-discovery learning) [23]. The experts are instructed to playtest the game in pairs, perform tasks together and uncover design issues indirectly via their discussions and explanations to each other. During playtesting, interaction between researchers and players (experts) is minimal; nevertheless, researchers keep notes, in a coding scheme. At the end of the playtesting session, players fill-in benchmarking questionnaires. We have applied field playtesting with experts' constructive interaction in the evaluation of two location-based MG-CH (that follow a common design approach) for two UNESCO intangible CH sites, and we report on several lessons learnt and insights.

978-1-6654-3886-5/21/$31.00 ©2021 IEEE

II. Background and Related Work

A. Expert Evaluation in Games

Expert-based evaluation in games has been largely approached as a problem of identifying heuristic guidelines that can be used in game inspections. One of the first attempts for heuristics about playability of games is presented in [6], who propose a list of 43 heuristics divided in four categories: game play, game story, mechanics, usability. In [31] a list of 32 playability heuristics to a heuristic evaluation is proposed according to the review of six games and 14 evaluators. Recently, in [34] a tool-based approach to usability evaluation of games is proposed, on the basis of a set of 237 heuristics to address that "*to date the most used guideline to evaluate games usability is still Nielsen's proposal, which is focused on generic software … most evaluations do not cover important aspects in games such as mobility, multiplayer interactions, enjoyability and playability.*"

From a practical aspect, inspection methods must rest on a short list of general heuristics, if at all, to be cost-effective. However, in the case of game design it seems an unattainable goal to produce a comprehensive, relevant and practical set of heuristics. Game design is characterized by several kinds of mechanics, dynamics, and aesthetics, while it is highly interactive, creative and often artful.

Addressing the lack of expert feedback in game evaluations can be accomplished with peer evaluation in games (i.e. by peer game designers or members of the team). In [30] types of peer feedback in game development based on interviews with game industry professionals in two game studios are presented. Three main types of feedback are identified: informal or drive-by or virtual, which is instant and opportunistic feedback usually provided through teleconferencing platforms; formal project feedback deriving from group meetings; team playtesting feedback, which is organized in a session where team members playtest the game and "*call-out issues as they encounter them*". We are interested in qualitative, yet formal feedback so that we can make use of the results in presentations to other stakeholders than the game design and development team.

In serious games, involving experts in the design and evaluation is standard procedure in domains like health, special education and rehabilitation [34]. But, in games for cultural heritage, the contribution of experts is mainly materialized in co-design approaches on initial and design phases of CH projects [4][5][33] and not for UX evaluation, despite several calls for this need (e.g. [22]). Local CH experts, depending on the context, may be museum curators, archaeologists, historians, architects, folklorists, musicologists, environmentalists, geologists, etc. Arguably, in digital CH, it is important to consider the views of experts about the cultural content of the game, as well as peers (game and interaction designers and technology developers). This is the approach taken in our work.

B. Playability and Playtesting

In game research, the focus of evaluation is often on playability [11], which has been defined as a concept with several dimensions. In [7] it is suggested that game playability evaluation falls into three main areas: (a) quality assurance which typically focuses on the quality of the game software and seeing that all the functionality is in place; (b) game usability testing; and (c) testing focusing on the game play. This is a similar view to that of [24] who defines playability as "*a term describing game quality. At minimum, it is formed by three components known as functionality, usability, and gameplay*." Functionality refers to the technical quality of the game, its smooth operation, without bugs or crashes, with short loading times, etc. Usability focuses on the game user interface so that the game is intuitive and easy to use. Gameplay refers to the rules of the game that create the mechanics, like the challenges and how to overcome them and the game plot.

Playtesting is a standard approach in game evaluation, which is described in [20] as "*formative evaluation conducted by external researchers during post-production, before final release, in order to identify adjustments that bring the game closer to the developer's intent*". According to [27], playtesting guides the design of a playful experience by generating detailed feedback to the development team about if and how the game fulfils the player experience goal and it is "*the kind of testing designers care about most*". According to [19] various methods are employed in playtesting in the game industry, including RITE (Rapid Iterative Testing and Evaluation, [16]) which employs observation and think-aloud techniques with the addition of an attending software engineer to rapidly alter the design based on the findings; as well as open-ended usability tasks, paper prototypes and empirical guideline documents. All these methods recruit players from the user community, not experts.

Playtesting is invaluable for the formative evaluation of MG-CH for two additional reasons: location and (cultural) content. Firstly, the location is connected to game challenges, which can be taken up only if the player is on the spot. Also, the player should be assisted or instructed to reach to specific locations, as well as guided about specific features within the correct location (e.g. to locate an architectural feature of a specific building). These issues can certainly be identified by end users; however, an expert-based review (from game and interaction designers) may predict and eliminate some of them beforehand. Additionally, issues of CH content validity, accuracy, and appropriateness must be addressed before user testing. These issues require expertise about local heritage and may not be predicted from heuristics or guidelines.

C. Constructive Interaction

We have adopted the approach of constructive interaction to engage the experts in dialog during playtesting. Constructive interaction is a research protocol with psychological foundations [17]. It has been originally proposed as a method of evaluation of HCI in [23] where an advantage of the method is that "*It is possible to allow the subjects to explore a problem and to develop the solution*"; this is suitable for the case of MG-CH that must be evaluated during exploration of both the game and the place.

Constructive interaction is not often employed in evaluation research, presumably because it is a formative method that produces qualitative findings specific to the system examined, while it also requires the recruitment of

more users comparatively to single-user testing methods. Nevertheless, it has proven to be very productive when compared to other usability evaluation methods: in [2] it is shown that children pairs identified more problems (both in total and of the most severe) than individual testers. In another analysis of several usability testing methods [12], it is reported that inexperienced evaluators identify almost double findings when cooperating with constructive interaction than when following any other method.

In the context of MG-CH, constructive interaction is appropriate for many reasons. The method resembles what people do when they visit cultural sites, i.e. they visit in pairs or small groups and learn how to play the game together, they seek and find answers to challenges while in-the-place, and ultimately learn about CH. In addition, constructive interaction affords ecological validity since that findings arise from players' interactions with minimal researcher intervention [9].

III. Field Playtesting with Experts' Constructive Interaction: Overview of the Method

A. Main Steps

The method of field playtesting with experts' constructive interaction must be employed when a functional prototype of the MG-CH has been developed but has not been yet tested in the location.

The method requires from experts to play the game in pairs, as they would do if they were visitors of a CH site. The experts-players are continuously observed by two researchers, who do not intervene in the process but record each player's actions, comments, questions (to one another).

During the process, researchers may further identify corrections that must be made, redesign ideas and other improvements. The researchers take notes about performance indicators, findings on playability and possible recommendations, preferably with digital devices that can also record videos of interactions and voice comments.

At the end of the playtesting, a discussion with experts can summarize findings and yield more general remarks; additionally, benchmarking questionnaires about the usability or aspects of the UX can be completed.

More specifically, we propose the following steps:

1) ***Introduction and setup.***

(a). Two experts are selected to cooperatively play the game. There must be at least one researcher per player, to be able to keep detailed notes on that player.

(b). The mobile game is setup on players' devices. Players start viewing the game, while researchers start keeping notes.

2) ***Playtesting.***

(a). Players cooperate and discuss about issues that attract their attention and interest.

(b). Researchers take notes on player performance, playability findings and ways to address them.

3) ***Wrap-up.***

(a). At the end of the game, players provide conclusive remarks to evaluators, who may in turn ask a few questions based on observations.

(b). Additionally, players fill-in benchmarking questionnaires about aspects of their experience.

B. Documentation and Notes-Taking

We propose the following taxonomy of issues about MG-CH, for which researchers can keep notes of player cooperative actions.

1) **Performance indicators**. Performance indicators are average values of player performance, which can be approximately anticipated from real users. Some performance indicators are generic and relevant to most MG-CH, while others heavily depend on each game. Based on our games, we propose the following performance indicators:

- Overall game time. This might be different in comparison to anticipated game time by end-users; even so, an approximation of overall game time is useful.
- Actual game time (for all challenges), i.e. without the time to walk from one mission or challenge to the next one.
- Time to complete a challenge (for each challenge), i.e. from the moment they realized they were on the spot until they provided an answer to the challenge.
- Distance covered. This indicator can report the actual distance covered by players including back and forth routes and wondering around.
- Experience points gained.
- Number of unsuccessful challenges. Although experts should provide lower values than end-users for unsuccessful challenges, this indicator can be considered for estimating challenge difficulty level.

2) **Playability findings**. In [24] a practical definition of playability is given, that "*contains only components that are designed into the game: gameplay, functionality and usability*". We further add two aspects specific for MG-CH: location context and CH content. Thus, we organize playability findings into the taxonomy:

- Gameplay findings are concerned with understanding game elements, rules, and dynamics (how these are updated and interact into a cohesive whole).
- Usability findings are about how easy it is for the players to make use of user interfaces and interactions in terms of performance and preference.
- Functionality findings concern the technical quality of the game, e.g. smooth operation, no bugs or crashes, short loading times, etc.
- Location context findings concern those related to information, indications and guidelines that help the user understand if she is at the right location or spot; or is heading rightly to it.
- CH Content findings concern the need to correct, add, create (new), or update CH content.

3) **Recommendations.** We propose keeping track of recommendations into the following types:

- Correction: an immediate, easy correction to the user interface or content of the game.
- Redesign (minor): that should require little design and development effort.
- Redesign (major/considerable): that should require considerable effort for design, development, and testing.

- Content creation: Need to correct or develop new CH content or considerably add to existing.
- Bug fix.

4) **Questionnaires for benchmarking**. We have asked players to fill-in the following standardized benchmarking questionnaires about usability and user experience:

- System Usability Scale (SUS, [13]), which consists of ten 5-point Likert statements. A score within [0, 100] can be computed as a quantitative indicator of usability.
- User Experience Questionnaire (UEQ, [29]), which consists of 26 pairs of terms of opposite meanings in a 7-point Likert scale, reflecting: attractiveness, classical usability aspects (efficiency, perspicuity, dependability) and user experience (originality, stimulation).

IV. Applying the Method in Two Mobile Games with a Common Design Approach

A. Emphasis on Storytelling and Exploratory Learning

The main goal of the MG-CH is to enable visitors discover and learn about intangible local heritage by walking through museums and settlements. The design of the games emphasizes storytelling and exploratory learning, which are interweaved with the main activity of the visit.

Storytelling is recognized as an important element of games in general [27], as well in CH games [14] [21]. It can give meaning to the goals, challenges and rules of the game as well as motivate the players. Storytelling creates an imaginary context in which the discovery takes place, combining real (e.g. historical information) and imaginary elements (e.g. roles and script). It enhances engagement and empathy with strangers / places / situations, as the player feels that he can participate and / or influence the story.

Exploratory learning is the natural process of first-time learning that emphasizes observation and exploration [26]. In MG-CH, exploratory learning is intertwined with visiting cultural heritage sites. More specifically, in both games:

- During gameplay, the visitor takes the role of helping a young character who is novice about local heritage and shares the goal of learning through undertaking missions located at the museum or in the settlement.
- Each mission is located at a specific area of the museum or the settlement and motivates the player to explore the place during the visit.
- Each mission comprises of challenges, that present specific questions (of multiple formats) to the player about points of interest like museum exhibits, signs, settlement monuments, buildings, and their architectural features.
- After the player completes a mission, a tool is earned and may be viewed and manipulated in augmented reality (AR). These tools are related to local craftmanship heritage; they are 3D models of real exhibits from within the museum collection.

B. Game 1: Exploring the Marble Town

1) ***Local context.*** Tinian marble craftsmanship (named from its origin: the island of Tinos, Greece) has been recognized globally, and since 2015 it is inscribed at the Representative List of Intangible Cultural Heritage of Humanity (UNESCO). Tinian marble-craftsmen acquire knowledge concerning marble and its properties, on a master-apprentice model through informal education, where the apprentice, after long and tough training, finally receives complete expertise. They have developed the relevant skills for the making of the tools used both in marble-crafting and marble-mining [8]. One of the most delegate villages of such rich and exceptional tradition of marble craftsmanship is the village of Pyrgos (Fig.1), where the involvement of the inhabitants with the marble craftsmanship has highlighted some of the greatest marble artists. Local heritage is presented at the Museum of Marble Crafts, which exhibits the equipment and explains the techniques used in order to mine and manipulate marble volumes, and then design, artistically create, and merchandise marble works, especially during the pre-industrial age, when Tinos was the most important center for the production and commerce of marble crafts in Greece.

2) ***Game plot.*** The goal of the game 'Exploring the marble town' is to connect the settlement of Pyrgos with the Museum of Marble Arts, to encourage players to visit both of them, and advance their knowledge about the heritage Tinian marble crafts. During the game the player uncovers the impact of marble arts in the life of the locals by exploring historical monuments, the quarries where local marble was sourced, visiting today's marble workshops in the settlement, and learning about works of art and tools in the museum.

Fig. 1. On the left: Panoramic view of the historic settlement of Pyrgos, Tinos island, Greece, and details on engravings and carvings that can be found throughout the settlement. On the right: Aspects of the museum of Marble crafts (work at the quarry, a craftsman's workshop, tools).

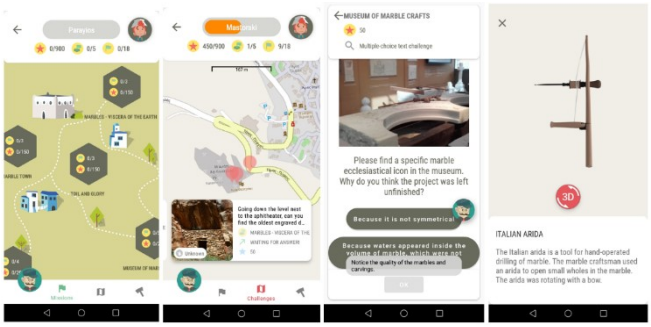

Fig. 2. Screen shots of the mobile game 'Exploring the Marble Town'. From left to right: (a) Missions view (graphical map); (b) View of a challenge (on a map background); (c) Answering a challenge; (d) Viewing a tool (earned after a mission).

The main plot puts the player in the shoes of a young marble crafter who is an apprentice to a senior local sculptor. The player is guided to points of interest where they are given a question to answer (challenge). Challenges are organized in groups (Missions) based on spatial and semantic criteria. Answers to challenges can be found in the physical context i.e. date carvings, building styling, marble signatures,

laboratory signs, etc. To answer a challenge, the player must interact directly with the site by observing, hearing, touching, thinking, and asking around for help. As players progress, they are awarded with experience points and traditional tools for their inventory (i.e. needles, hammers, etc.), and level-up to evolve as senior sculptors in marble crafts. At the end of the game, the players have completed a playful tour around the settlement of Pyrgos and the museum of marble crafts, and they have explored the impact of marble in the local life; additionally, they have 'climbed the ladder' from being an apprentice to being a senior marble craftsman. Screen shots of the mobile game are shown in Fig. 2.

C. Game 2: The People's Machine

1) ***Local context.*** At the heart of the Mediterranean diet (inscribed at the UNESCO Representative List of Intangible Cultural Heritage of Humanity) is the use of olive oil. The island of Lesvos in Greece is renowned for its olive and olive-oil production that dates back to antiquity. At the end of the 19th century, the economic demands and the industrial evolution lay the foundation for a civil society. The citizens of the settlement of Agia Paraskevi committed to the creation of a community's olive mill, the 'people's machine', that would escape the commission and monopoly tactics of the corresponding private olive mills in the area [25]. Its proceeds would be used for the construction of the village's schools and other public benefit purposes. The 'people's machine' has now turned into a museum, the Museum of Industrial Olive-Oil Production of Lesvos (Fig. 3), which seeks to promote and safeguard the industrial heritage of the place and to integrate it into the broader architectural, social, and cultural context of the time of its prosperous period.

2) ***Game plot.*** The main goal of the game 'The people's machine' is to create awareness about how a small community with a vision in the island of Lesvos, managed to achieve economic, social, and educational uplift by producing olive oil. During the game, the player uncovers elements of the true story of the construction of the local communal olive oil mill (now museum of industrial olive oil production) by exploring the museum and visiting the nearby settlement to discover parts of the story in buildings, signs and monuments.

The main plot puts the player in the role of a young man living in the era in which the idea of a communal olive-oil production unit was conceived. According to the story, every resident of the village contributed in their own way to the realization of the common vision, through voluntary work, spreading the idea via 'word of mouth', with financial support if they could afford it, and so on. Each mission consists of challenges related to a specific point of interest around the settlement or the museum. The player is asked to discover and collect historical data by observing the environment at those points of interest, through which the narrative progresses. By completing missions, the player is being awarded with experience points and traditional tools related to olive-oil production i.e. wrenches, shovels, olive-oil storage jars, etc. These tools can be found in the player's inventory throughout the game. At the end of the game, the players have completed a playful tour around the settlement and the museum, and have appreciated the impact of olive oil in the local life; they also have 'climbed the ladder' from being an apprentice to being a master olive worker in the mill. Screen shots illustrating the mobile game are shown in Figure 4.

Fig. 3. On the left: Panoramic view of the settlement of Agia Paraskevi, Lesvos island, Greece. On the right: Aspect of the interior of the museum of industrial oil production.

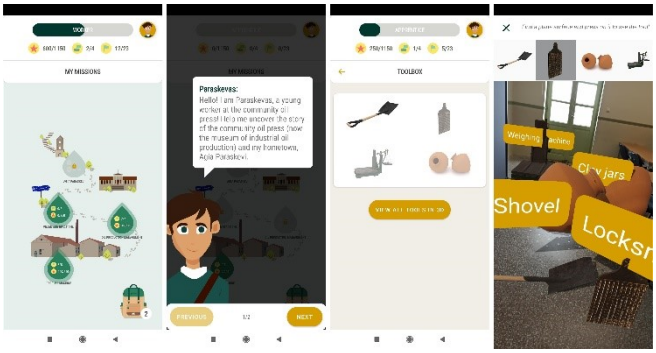

Fig. 4. Screen shots of the mobile game 'The people's machine'. From left to right: (a) View of the missions (last mission locked); (b) View of the user character, introducing himself; (c) View of the Toolbox; (d) Viewing the tools earned in AR.

D. Expert participants

Ten experts were recruited in playtesting of game 1, average age 41 years, four women. Four of them were CH experts: the museum director, two museum staff and a local heritage expert. Other experts were: two game designers and developers, two interaction designers, one graphics and 3D content developer, and one from IT (Information Technology) and CH project management.

Fourteen experts were recruited in playtesting for game 2, average age 39 years, seven women. Six of them were CH experts: the museum curator, the museum director, two museum staff and two local heritage experts. Other experts were: four game designers and developers, two interaction designers, one graphics and 3D content developer, and one from IT (Information Technology) and CH project management.

Both playtesting sessions took place during two-day project meetings at the museums of the sites. During the playtesting sessions, four HCI researchers in total observed all players (Fig. 5). For each playtesting session, a pair of expert players was formed as well as with a pair of researchers. Each researcher was focusing on the movements, comments, and behavior of a particular player. After each researcher recorded findings on his/her own, they cooperatively processed and characterized them.

E. Results

1) ***Performance indicators.*** As shown in Table 1, Game 1 lasted more than game 2 since that most challenges (13 from 18) were to be discovered outdoors (in the settlement), which also increased the average distance covered by each pair of players. For Game 2, most challenges (9 from 15)

were located within the museum and its surrounding space; thus, gameplay time and distance covered were decreased.

Both games were played at players' pace, meaning that they all continuously interweaved gameplay with walking around the museum or the settlement to see the sites and discuss about them, as would players normally do in a visit. For both games, all player pairs undertook all challenges with a high level of success. The researchers retrospectively assessed the difficulty of the challenges with characterizations easy, medium, hard, on the basis of discussion with participants at the wrap-up session.

All data about performance indicators could not be safely estimated beforehand. These are valuable data that can be used to further enhance the UX of the games by introducing them in the gameplay, e.g. to the reward system or to add indications or warnings.

Table 1. Performance indicators

Average values	Game 1	Game 2
Game completion time	107.6 min	66 min
--Gameplay time	54.7 min	40 min
--Time in-between gameplay	52.9 min	26 min
Distance covered	3.3 km	1.4 km
Challenge time to complete	3 min	2.7 min
Experience points gained	790 (max 900)	670 (max 750)
Challenges not answered correctly	2.9 (total 18)	3 (total 15)
Challenges' difficulty	4 hard; 9 medium; 5 easy	3 hard; 8 medium; 4 easy

2) ***Playability****.* We discuss playability findings (Fig. 6) for general aspects of the game (onboarding screens, mission information, user profile, etc.) as well as for challenge-specific aspects (questions of multiple formats, in specific locations), in the taxonomy proposed previously (section III.B), including: gameplay, usability, functionality, location context, CH content.

Firstly, we observe there is a common pattern of playability findings for both games; that was anticipated since that the games have the same system and gameplay design. We had also anticipated to find more CH content issues in Game 2 (because we had less time to work on these before playtesting compared to game 1), which was also confirmed.

If we first look at general playability findings (left columns of Fig. 6), we can observe that most issues were about usability. Examples of usability issues include: *"It is not perceived by users that the helper character icon can be clicked"; "Missions accomplished are not highlighted onto the mission map",* etc. We found that usability issues were mostly of a general nature involving navigation, user guidance and help, showing/notifying the dynamics of the game. It is no surprise that usability is important, but clearly it is not the only concern of evaluation in games, especially when we examine the detailed game play (answering to challenges), where there are not many issues; we found only a few challenge-specific usability issues, like that *"When filling in a response, sometimes the pop-up keyboard hides the input control."*

In addition, a considerable number of general playability findings are about functionality of the games. For example, *"The app crashes in Android versions close to the minimum version", "The GPS is not precise, when the player is inside the museum", "The player must be given the option to download offline maps",* etc. Another considerable number of issues are about gameplay, for example: *"Missions (and challenges) must be illustrated in the (suggested) order in which they can be undertaken", "Messages about user success/failure must first reveal the correct answer (consistent structure of messages)".* Functionality and gameplay issues are also more intense when players access more general functions like missions, rewards, maps, profile, etc.

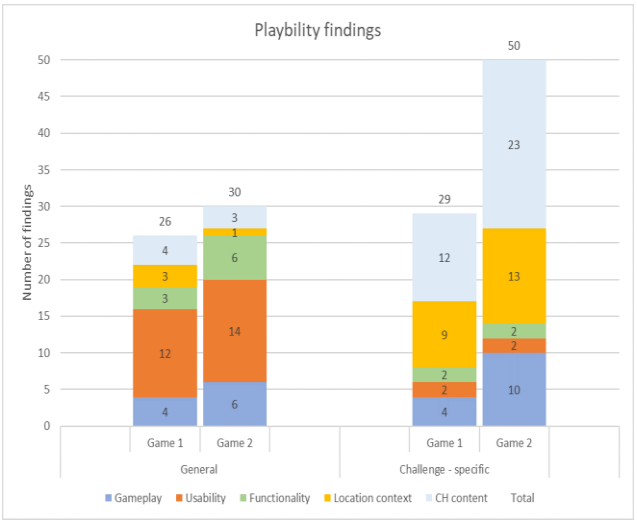

Fig. 5. Playability findings (unique, per game and aspect).

The detailed review of general findings from two games revealed that a thorough set of findings was identified already from the first game – only a few new issues were identified in the playtesting session of the second game. This is a supportive finding about the breadth of coverage of findings for the proposed evaluation method. Furthermore, we suggest that it is important to examine all tasks and aspects of the game, in contrast to some usability testing approaches that pick some user tasks only.

When we look at challenge-specific playability findings (right columns of Fig. 6), issues about CH content and location context prevail. Regarding CH content, findings include several corrections to texts and photographs. More importantly, experts marked some challenges that did not convey an important takeaway about CH; these were replaced or reworked, and this actually occurred in-situ. Regarding location context, findings include several corrections on guiding users where to navigate or look to find answers. In particular, the GPS user location did not suffice; the players required references to landmarks and other signs. Examples of findings about location context include: *"Some users are not sure if they are at the right spot in order to start looking for the answer (they wander around unnecessarily)", "Some users cannot locate the answer, despite on the spot (they need to be provided with a guideline, e.g. read the sign)".* Thus, it was invaluable to employ CH professionals and local experts in a playtesting approach, which enabled them to examine every single detail of the game.

3) ***Recommendations****.*

Most of general recommendations *(Fig. 7)* fall into minor redesign actions required, like for example *"Add short animation to the character icon to denote that he is a helper to the user and that it is clickable"; "To clearly highlight all completed missions on the map and during browsing",* etc. There were also some redesign issues that will require considerable effort, like that *"To add more onboarding screens about characters of the game", "To redesign the map of missions to put them in exact order with respect to their location"*. These findings confirm the value of conducting playtesting sessions in the field, since that it is obvious that they are easy to identify when playing at the location.

Most recommendations about the challenges of the game are corrections that must be made, mainly to CH content, for example *"Fix naming of the player levels", "Ensure consistent English terminology with the signs of the museum", "Add explanation to reward message (challenge-specific)",* etc. Another major area of recommendations is about the need to create new CH content. Since we were in the field, we rapidly drafted the required information, took required photos, etc. These findings further confirm that involving local CH experts in playtesting is invaluable: they are the only expert group that can confirm the accuracy and validity of CH content as well as they can easily provide additions and alternatives to incomplete or invalid content.

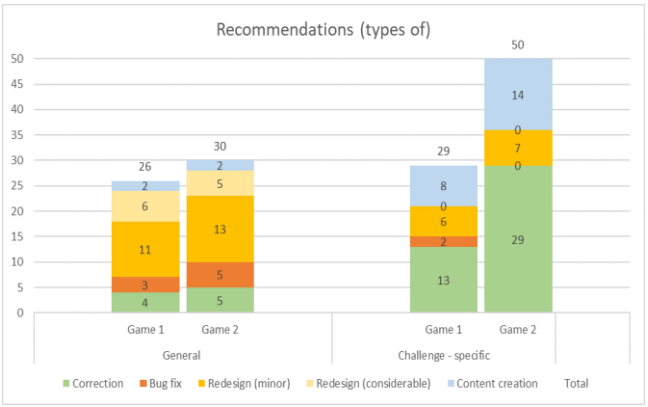

Fig. 6. Recommendations (types of, per game).

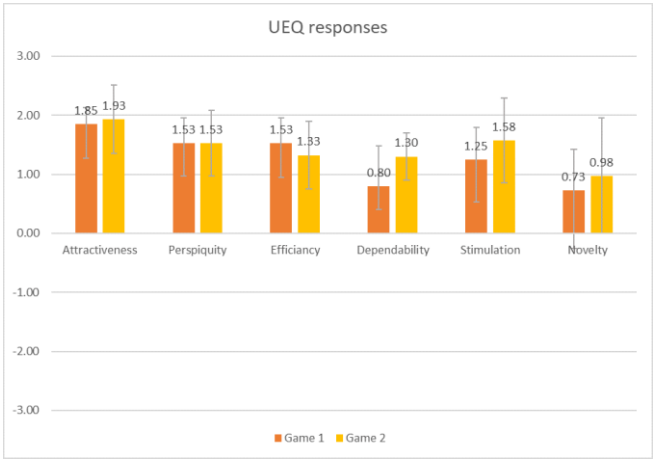

Fig. 7. UEQ responses.

1) ***Questionnaire responses****.* At the end of each session, we asked participants to fill-in benchmarking questionnaires about perceived usability and UX.

Perceived usability was measured with the SUS questionnaire [13]. The SUS scores are satisfactory and close to each other (81.8 for game 1, 83.4 for game 2), although they imply room for improvement. The SUS score is very satisfactory when above 80, it is fairly satisfactory when between 60 and 80, and not satisfactory when below 60 [1].

Perceived UX was measured with the UEQ questionnaire [29]. The results are depicted in Figure 9. For both games, no significant differences are observed in all dimensions (which was anticipated). All dimensions are positively ranked and range in [0.73, 1.93]. According to [28] "*the standard interpretation of the scale means is that values between −0.8 and 0.8 represent a neural evaluation of the corresponding scale, values > 0.8 represent a positive evaluation and values < −0.8 represent a negative evaluation*."

V. Discussion and Conclusions

This work makes a purposeful synthesis of existing HCI evaluation methods and proposes a new, formative method of field playtesting with experts' constructive interaction for application in MG-CH. The proposed method:

- It is a design review method (rather than an inspection), since that it does not rest on heuristics or guidelines, but on expert opinion.
- It is an empirical method, which unfolds as experts experience the game, in contrast to typical design reviews or juries that examine presentations or demonstrations of artefacts or systems. In comparison to heuristic evaluation where 3-5 double experts are proposed [17], a larger pool of experts is assumed. We do not propose employing game design experts only (as in [10]); we consider essential the participation of CH professionals, interaction designers and software developers.
- It requires playtesting, a standard approach to game evaluation that emphasizes playability, i.e. a composite concept that includes (at least) aspects of gameplay, usability, and functionality [24].
- It requires from experts to actively play the game following the approach of constructive interaction, in contrast to other expert-based approaches that rest on inspection or review.

We have analyzed the rationale of the method and we have presented a case study of two games with same system and gameplay design for which the method yields consistent and rich results. The main conclusions from these studies include:

- The method yields several performance indicators that can be exploited in gameplay design and may not be safely estimated beforehand. We have presented some generic performance indicators that may be relevant to other MG-CH. Depending on the game at hand, more performance indicators can be identified.
- The method is very productive in findings about several dimensions of playability. We have presented several qualitative findings classified into the dimensions of gameplay, usability, functionality, location context and cultural content.
- The method is very productive in recommendations for redesign. Experts and evaluators work in-context, which

enables them to readily identify actions for improvement, which are rapidly produced and confirmed.

- The method affords readiness to generate new cultural content. It should be easy for the evaluators to work with experts in-situ to create new content (in the case of our games this included: questions, hints, guidelines to locate place, GPS points, photos, etc.).
- The method increases active contribution of CH experts in evaluation of MG-CH. It is important for evaluators to involve experts in ways that enable them to provide specific comments rather than general guidelines.
- The evaluation process was fun for all participants.

The method of field playtesting with experts' constructive interaction is original to the extent that it synthesizes the ideas of playtesting in the field, recruiting experts in CH and design, and applying co-discovery learning in HCI evaluation practice of iterative development of mobile games for cultural heritage. This work may contribute and complement to co-design approaches in CH in a novel manner that capitalizes on HCI evaluation methods. It also offers a practical way to involve CH and design experts into the process of playtesting of MG-CH. We envisage that the proposed method can be adopted and applied to other situations of MG-CH.

VI. Acknowledgements.

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under call RESEARCH–CREATE–INNOVATE (T1EDK-15171).

References

[1] W. Albert T. Tullis. 2013. Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes.

[2] B.S. Als J.J. Jensen M.B. Skov. 2005. Comparison of think-aloud and constructive interaction in usability testing with children. In Proc. of the 2005 conference on Interaction design and children (pp. 9-16). https://doi.org/10.1145/1109540.1109542

[3] N. Avouris C. Sintoris N. Yiannoutsou. 2018. Design guidelines for location-based mobile games for learning. In Proc. of the 17th ACM Conference on Interaction Design and Children (pp. 741-744). https://doi.org/10.1145/3202185.3205871

[4] L. Ciolfi G. Avram, L. Maye, N. Dulake, M.T. Marshall D. van Dijk F. McDermott. 2016. Articulating co-design in museums: Reflections on two participatory processes. In Proc. of 19th ACM Conference on CSCW & Social Computing, 2016, pp. 13-25.

[5] V. Cesário A. Coelho V. Nisi. 2018. Co-designing gaming experiences for museums with teenagers. In Interactivity, Game Creation, Design, Learning, and Innovation; Springer, Cham, 2018, pp. 38-47.

[6] H. Desurvire M. Caplan J.A. Toth. 2004. Using heuristics to evaluate the playability of games. In CHI'04 extended abstracts on Human Factors in Computing Systems https://doi.org/10.1145/985921.986102

[7] M.P. Eladhari E.M.I. Ollila. 2012. Design for research results: experimental prototyping and play testing. Simulation & Gaming, 43(3), 391-412. https://doi.org/10.1177/1046878111434255

[8] A. Florakis. 2009. Museum of Marble Crafts, Guidebook. Piraeus Bank Group Cultural Foundation, Athens. ISBN: 9789602441404

[9] H. Kahler F. Kensing M. Muller. 2000. Methods & tools: constructive interaction and collaborative work: introducing a method for testing collaborative systems. ACM interactions, 7(3), 27-34. https://doi.org/10.1145/334216.334226

[10] H. Korhonen. 2010. Comparison of Playtesting and Expert Review Methods in Mobile Game Evaluation. In Proc. of the International Conference on Fun and Games 2010, Leuven, Belgium, 18-27. New York, NY, USA: ACM. doi:10.1145/1823818.1823820

[11] H. Korhonen. 2016. Evaluating playability of mobile games with expert review. https://www.semanticscholar.org/paper/Evaluating-Playability-of-Mobile-Games-with-the-Korhonen/80e976e4571bed7a20bad52e209e365fd5532557

[12] P. Koutsabasis. T. Spyrou, J. Darzentas. 2007. Evaluating usability evaluation methods: criteria, method and a case study. In: Jacko J.A. (eds) Human-Computer Interaction. Interaction Design and Usability. LNCS 4550. Springer. https://doi.org/10.1007/978-3-540-73105-4_63

[13] J.R. Lewis. 2018. The system usability scale: past, present, and future. Int. J. of Human–Computer Interaction, 34(7), 577-590. https://doi.org/10.1080/10447318.2018.1455307

[14] I. Malegiannaki. T. Daradoumis. 2017. Analyzing the educational design, use and effect of spatial games for cultural heritage: A literature review. Computers & Education, 108, 1-10. https://doi.org/10.1016/j.compedu.2017.01.007

[15] L.A. Maye F. McDermott L. Ciolfi G. Avram. 2014. Interactive exhibitions design: what can we learn from cultural heritage professionals? In Proc. of the 8th NordiCHI, 598-607. ACM. https://doi.org/10.1145/2639189.2639259

[16] M.C. Medlock D. Wixon M. McGee D. Welsh. 2005. The Rapid Iterative Test and Evaluation method: Better products in less time. In Cost-justifying usability (pp. 489-517). Morgan Kaufmann. https://doi.org/10.1016/B978-012095811-5/50017-1

[17] N. Miyake. 1986. Constructive interaction and the iterative process of understanding. Cognitive science, 10(2), 151-177.

[18] J. Nielsen. 1992. Finding usability problems through heuristic evaluation. In Proc. of the SIGCHI Conference on Human factors in computing systems. https://doi.org/10.1145/142750.142834

[19] P. Mirza-Babaei N. Moosajee B. Drenikow. 2016. Playtesting for indie studios. In Proc. of the 20th International Academic Mindtrek Conference. https://doi.org/10.1145/2994310.2994364

[20] P. Mirza-Babaei S. Stahlke G. Wallner A. Nova. 2020. A Postmortem on Playtesting: Exploring the Impact of Playtesting on the Critical Reception of Video Games. In Proc. of CHI Conference on Human Factors in Comp. Systems. https://doi.org/10.1145/3313831.3376831

[21] M. Mortara E. Catalanoa F. Bellotti G. Fiuccic M. Houry-Panchetti P. Petridis. 2014. Learning cultural heritage by serious games. J. of Cult. Heritage, 15(3), 318-325. https://doi.org/10.1016/j.culher.2013.04.004

[22] V. Nikolakopoulou P. Koutsabasis. 2020. Methods and Practices for Assessing the User Experience of Interactive Systems for Cultural Heritage. In Applying Innovative Technologies in Heritage Science (pp. 171-208). https://doi.org/10.4018/978-1-7998-2871-6.ch009

[23] C. O'Mailey S. Draper M.S. Riley. 1984. Constructive interaction: a method for studying user-computer-user interaction. In IFIP INTERACT'84 1st Int. Conf. on Human-Computer Interaction.

[24] J. Paavilainen. 2017. Playability: A Game-Centric Definition. CHI PLAY '17 Extended Abstracts. October 2017. Pages 487–494. https://doi.org/10.1145/3130859.3131306

[25] C. Paraskevaides. 1991. The old Agia Paraskevi of Lesvos. Agia Paraskevi Community Cultural Center. ISBN: 960-220-113-4

[26] J. Rieman. 1996. A field study of exploratory learning strategies. ACM TOCHI, 3(3), 189-218. https://doi.org/10.1145/234526.234527

[27] J. Schell. 2015. The Art of Game Design: A Book of Lenses. CRC Press. ISBN: 978-0123694966

[28] M. Schrepp. 2018. User Experience Questionnaire Handbook; Accessed at: https://www.ueq-online.org/ Material/Handbook.pdf

[29] M. Schrepp A. Hinderks J. Thomaschewski. 2017. Construction of a benchmark for the User Experience Questionnaire (UEQ). Int. J. of Inter. Multim. and AI, 4(40-44).

[30] J. Seering R. Mayol E. Harpstead T. Chen A. Cook J. Hammer. 2019. Peer Feedback Processes in the Game Industry. In Proc. of CHIPlay (pp. 427-438). https://doi.org/10.1145/3311350.3347176

[31] S. Soomro W. Fatimah W. Ahmad S. Sulaiman. 2013. Evaluation of mobile games using playability heuristics. In: Advances in Visual Informatics (IVIC 2013). LNCS, vol 8237. Springer, Cham. https://doi.org/10.1007/978-3-319-02958-0_25

[32] UNESCO, World Heritage Centre, https://whc.unesco.org/

[33] S. Vosinakis, V. Nikolakopoulou, M. Stavrakis, L. Fragkedis, P. Chatzigrigoriou, & P. Koutsabasis. 2020. Co-Design of a Playful Mixed Reality Installation: An Interactive Crane in the Museum of Marble Crafts. Heritage, 3(4), 1496-1519.

[34] R. Yáñez-Gómez D. Cascado-Caballero J.L. Sevillano. 2017. Academic methods for usability evaluation of serious games: a systematic review. Multim. Tools and Appl., 76(4), 5755-5784. https://doi.org/10.1007/s11042-016-3845-9.