Methodology for Playtesting Computer Games

A Mixed Method Approach

Mary Magee Quinn, Carl Symborski, Meg Barton
Science Applications International Corporation
Arlington, VA, USA

James Korris, Travis Falstad, Stephanie Granato
Creative Technologies Inc.
Los Angeles, CA, USA

Abstract— **This paper describes a mixed method approach to playtesting a serious computer game designed to identify and mitigate cognitive bias. This approach incorporated in-depth data collection through the use of recordings of the screen and voice of the player while testing, a detailed survey, and a follow-up focus group. The data collection methodology served to provide detailed feedback from playtesters that was analyzed and used to make continuous and timely changes to the game.**

Keywords—playtesting, usability, playability, video games

I. Introduction

Computer games are intended to present the end users with challenges, engaging players through their stimulating content. However, one of the aims of the game developer is to eliminate unintended challenges such as those caused by mechanical or usability problems, while ensuring that gameplay is as fluent and engaging as possible, rendering the game easy to learn and to master [1], [2]. To test for unintended challenges and to measure engagement, game developers use playtesters to play and provide feedback about the game [2]-[4]. In the best cases, playtesting is an iterative process in which testers with similar characteristics to the anticipated end user (e.g., age, education level, professional similarities, gaming experience) test initial and subsequent builds of the game and provide feedback to the game designers, which is then incorporated into the game design [3], [5].

There is an expansive literature base on usability testing, and many scattered resources on heuristic approaches to playtesting and game assessment. However, there are few brief, comprehensive "how to" guidelines detailing best practices for playtesting computer games in the literature base. This paper[1] presents an overview of a playtesting method for computer games, informed by the relevant literature and shown to be effective when put into practice, that can be employed by a development team in a time- and cost-effective manner. It incorporates both qualitative and quantitative data elements, and both informal/internal playtesting and formal playtesting with recruited players.

In the context of a "serious game", one with a need to inform as well as engage, the audience may reflect a broad spectrum of player experience and proficiency. In this instance, early and consistent feedback on the game's usability is of the essence, with playtesting findings significantly influencing many aspects of the game. The user experience, including the user interface, player navigation, hinting/tutorial system, along with pacing and timing, are the most likely to improve with timely input from a playtesting effort.

The playtesting method used in this paper was applied as part of the development of a serious computer game designed to teach the recognition and mitigation of cognitive biases. Since "issues in usability can drastically impact user experience and thus the learning outcomes associated with serious games" [2], the playtesting protocol described below was developed to minimize usability/playability issues and maximize the game's effectiveness. While it was developed to test a serious game, this playtesting protocol is generalizable to any type of computer game. What follows is a description of that protocol, along with a discussion of associated results and implications.

II. Informal Playtesting Method

This playtesting method includes two facets: informal and formal playtesting protocols. Informal playtesting was primarily conducted by two to four project team members at any given time. Internal playtesting team members tested each build of the game as it became available. These internal testers were all educated as to the purpose and intricacies of the game, and carefully tested all possible features and ferreted out bugs. This was similar to the "game breaking" method, in which the full functionality and limits of the game are tested, as described in [2].

Informal playtesting began as new builds of the game were released to the internal playtesting team, after preliminary testing by the game company to ensure that the build ran properly and that there were no major flaws. The internal playtesting team consisted, in total, of three males and two females, with a range of ages and varying game experience, from just a few months to more than 20 years. The internal playtesters tested the game through several run-throughs, making detailed notes as they played. These notes were aggregated daily across the internal playtesting team members,

[1] This work supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

formatted into a document with action items organized by priority, and released to the game development team. The list of action items was maintained throughout the development cycle, with items being removed from the list once they were fixed and additional action items being added as necessary. In addition, the game developers transferred every issue into a bug database that encompassed the results of their own internal testing as well. Weekly conference calls between the internal playtesting team and the game development team were held to ensure that all feedback was clear and manageable. This iterative process was used on every new game build.

By using internal playtesters knowledgeable about the game to perform debugging tests, formal playtesting with recruited test users could instead focus on a holistic evaluation of the game experience.

III. FORMAL PLAYTESTING METHOD

Formal playtesting was conducted with playtesting volunteers external to the project team. The process entailed uniformly following a playtesting protocol that included an introductory session, a play session with affordances for data collection, a post-playtest heuristic questionnaire, and a focus group debriefing session. The results of the formal playtesting were analyzed and reported to the game development team in an iterative cycle.

A. Heuristic Evaluation

Heuristic evaluation is a method of game evaluation in which a set of guidelines is used to “rapidly identify common issues in game design,” [6], [7]. Heuristics are usually employed by expert game evaluators, and are generally considered an efficient, low-cost alternative to user-testing [6], [7]. In this protocol, however, heuristics served as the foundation for the user-testing procedure, from the data collection tools to the analysis of the data.

While there are a number of available heuristics for game evaluation [7], the Heuristic Evaluation for Playability (HEP) has been demonstrated to be effective and is fairly comprehensive. The HEP includes four major heuristic categories: gameplay, mechanics, game story, and usability. The first heuristic, gameplay, was designed to determine the user’s perceptions, attitudes, and opinions about interacting with the game. The game mechanics heuristic was used to determine how the rules and functions of the game impacted the user experience. The game story heuristic was used to determine the user’s perceptions, attitudes, and opinions about the underlying story behind the game. Finally, the usability heuristic was used to determine whether the player experience (e.g., screens, displays, menus, controls) matched the design intent [8]. Project team members collaboratively developed specific research questions for each of the four heuristic categories, informed by [5], [7], and [8]. Table I outlines our four heuristic objectives in more detail.

TABLE I. GAME EVALUATION HEURISTICS

Heuristic	**Description**
Gameplay	Does the game provide clear goals for the user?
	Does the player see the progress in the game?
	Does the player feel in control of the game?
	Are the challenge, strategy, and pace balanced?
	Was the first-time experience encouraging?
Mechanics	Are the game mechanics consistent throughout the game?
	Are the controls easy to learn?
	Does the navigation system support the ease of gameplay?
	Is it easy to explore the playfield?
Game Story	Is the game story meaningful?
	Are there repetitive or boring tasks?
	Does the player have the opportunity to express him/herself?
	Does the player relate to the characters?
	Did the gameplay make sense with the story?
Usability	Is the user interface consistent throughout the game?
	Is the user interface similar to other games the user has experienced?
	Is the feedback to the user from the game adequate?
	Is all information that the user needs displayed clearly when the user needs it?
	Is the screen layout efficient?
	Is the screen layout visually appealing?
	Does the visual appearance support the playing of the game?
	Do the audio effects support the playing of the game?

B. Formal Playtesters

It is important for the players in a playtesting study to mirror the intended end user population [2], [5]. For this study, the serious game under development was intended for use by entry level employees who would require training on cognitive bias mitigation. Accordingly, formal playtester recruitment focused on young adults with similar levels of education and digital nativity as would be seen in the target population. Twenty-one representatives of the sponsoring customer, as well as subject matter experts (SMEs) in the training of cognitive bias recognition and mitigation, participated in formal playtesting to provide their expert feedback.

A second concern when it comes to playtester selection is the number of players required to effectively expose the major problems of the game [9]. While some experts agree that a well-planned playtest with four to five players can expose up to 80% of usability difficulties and that that 80% will likely represent the major problems for any build [10], others suggest that five playtesters is far too few [11]. Considering the rapid turn-over in game builds, however, it often does not make sense to test more than four to five people, as a new build of the game may be available after the testing of those five participants has been completed. Additionally, it is more cost-effective to run fewer playtesters and, in the project team's observations, after four to six players had completed a playtesting cycle, many of their comments were redundant and covered comprehensive feedback on that build.

In this study, multiple cycles of formal playtesting were conducted, generally with two to six players per build. This resulted in a total of 35 non-customer identified formal playtesters for 12 builds. In addition, four playtesters who played initial builds in May 2012 were asked to play the final build in March 2013 to enable us to compare feedback and determine whether improvements were noticeable.

C. Setting

Formal playtesting took place in a specially designed lab, which was set up to accommodate two playtesters simultaneously. Desks were arranged so that playtesters could view each other's computer screens to encourage communication. The lab also contained a rear-shoulder view mounted camera to allow observers to stream video of the lab without being physically present during playtesting. Frequently in playtesting sessions, the researcher hovers over the player and asks probing questions about the gameplay experience as the player progresses through the game [8]. This creates an unnatural game experience in which the player is distracted and unable to become fully immersed in the game experience. To avoid this, the team implemented a remote monitoring system that would alert researchers when a playtester was finished and/or if (s)he needed help; otherwise, playtesters were left to test in an environment similar to the intended end user play environment.

D. Materials

The research objectives and questions informed the selection and design of data collection tools. Several data collection tools were used, including a demographics form, a video/audio recorded gameplay session, a post-playtesting questionnaire, and post-playtesting focus group questions. Following is a description of each of the data collection tools.

1) Demographics: It is important to understand whether the playtesters replicate the characteristics of the intended end users of the game, which requires some demographic information. Demographic information may also be useful in more detailed analysis of playtester comments [2]; for example, analysis may reveal that novice gameplayers have great difficulty with navigation, whereas more advanced gameplayers have less difficulty. This may suggest to the developer that the addition of a navigation tutorial at the opening of the game may aid novice players with this unintended challenge, along with the incorporation of alternate ways to move the character.

The developed demographics form included questions to determine user gender, age, race, education level, and video gaming experience. Demographic data were summarized using descriptive statistics and were used in analysis of the playtesting data.

2) Video/Audio Recorded Gameplay Session: Gameplay recordings of the playtesting sessions were captured using a commercially available screen capture software that recorded both the player's screen and his or her speech via a headset microphone. Most playtesters agreed to have their voices and screens recorded and signed an authorization for video/audio recording form.

Playtesters were encouraged to "think aloud" while playing the game and to discuss any interesting, frustrating or confusing experiences with his or her playtesting partner. This arrangement incorporates two well-established playtesting techniques: the think-aloud protocol [12] and paired-user testing [9]. The think-aloud protocol, as its name implies, involves the playtester speaking aloud any thoughts, actions, or feelings he or she has while playing the game, thereby providing "live" feedback on game content as it is occurring [12]. This method is valuable in that specific, small-scale comments are reported in the moment, rather than forgotten by the time of post-playtesting data collection, and players do not have to stop and write comments and suggestions while playing. Additionally, the playtesting video/audio recording provided a wealth of detailed data to the game development team. There are, however, several perceived drawbacks to this method. Some suggest that it alters the flow of the game experience, while others suggest that it places too much cognitive load on the playtester, who then cannot concentrate fully on the gameplay experience. Furthermore, the data-rich video/audio recordings produced may take an extensive amount of time to analyze [2]. These drawbacks were ameliorated by the playtesting protocol design, and did not appear to cause any issues. Though the playtesters did receive initial encouragement to speak their thoughts aloud, they received no further prompts, thereby assuring that each participant only spoke as much as was natural and comfortable for that individual within the flow of gameplay. To minimize the amount of time required to analyze video/audio recordings of playtesting sessions, all comments were simply transcribed, and later added to the pool of formal playtesting data that was then further analyzed. No complicated behavioral analysis measures requiring time-consuming video scoring techniques [4], [9], were conducted. In paired-user testing, two playtesters may cooperate to complete a task on one computer; in this case, each playtester was at his or her own computer station, but both participants played through the game at the same time, and were encouraged to talk to one another. This enriched the quality of the think-aloud data, because playtesters feed off of one another's comments. This setup also relieved some of the awkwardness that playtesters may experience when speaking aloud while alone [9]. These video/audio recordings were reviewed by researchers to document user experience and, in a few cases, provided detailed information needed to isolate the causes of bugs that were identified during playtesting.

3) Post-playtesting Questionnaire: A post-playtesting questionnaire was developed collaboratively between the researchers and game developers and was piloted by internal playtesters to further refine the questions before being deployed. The questionnaire was divided into the four heuristic areas detailed above. The survey questions corresponded to the types of questions in Table I, but were designed to be more specific to the game design, mechanics, story, and NPC characters of the serious game being tested (e.g., “The audio effects in the apartment were realistic.”). Responses to statements about the game experience were measured using a five-point Likert scale (i.e., strongly agree, agree, uncertain, disagree, and strongly disagree), generating quantitative data, and a section for playtester-provided comments, generating qualitative data.

Since the questionnaire was fairly lengthy, several “red herring” questions were added to help determine whether respondents were carefully processing and responding to the questions. For example, one item read: “It was easy to navigate my character around the basement.” However, there was no basement scene in the game. The inclusion of obvious red herring questions on questionnaires has been demonstrated to be effective at “induc[ing] survey respondents to provide higher quality data at the outset,” as the presence of red herring questions indicates that the questionnaire developers place a high value on test-taker effort and data accuracy [13].

4) Focus Group Questions: Since the questionnaire and the audio/video recording rendered specific data, the focus group was designed to initiate a less structured, general discussion of the gaming experience. Focus groups are frequently used in game testing to gather in-depth feedback from a subset of the intended end user population. Conducting a focus group with several playtesters creates the opportunity for a discussion that builds on itself, with each individual expanding on the others’ feedback, to provide a wealth of rich qualitative information [10]. Three basic questions led the focus group discussion:

- What did you like about the game?
- What frustrated or confused you about the game?
- What do you consider the priority issue for the game developer to address?

E. Design/Playtesting Process

To encourage standardization, a playtesting orientation script was developed and followed for each of the playtesting sessions. It is described in the following sections.

1) Orientation: Playtesters arrived at the playtesting site in groups of two and were escorted by staff members to the playtesting room. To begin, playtesting researchers and testers would briefly introduce themselves. A researcher would describe the vision of the game and the current build. The lead researcher would then describe each of the heuristic areas of interest: gameplay, mechanics, game story, and usability.

The lead researcher continued by reinforcing what the playtesting exercise was *not* designed to test. Informal playtesters had been trained to seek out bugs, and while a formal playtester’s discovery of a bug would be valuable, the purpose of formal playtesting was not debugging but was instead geared toward gaining a naïve player’s holistic assessment of the game content. Further, players were reassured that their gaming skills were not being assessed; in fact, any issues that they encountered were the “fault” of the game, and could be addressed by the developers accordingly. Finally, though the game in question being tested was a serious game, the project team opted to conduct testing of the game’s effectiveness at teaching cognitive bias identification and mitigation separately, and to focus solely on gameplay, mechanics, game story, and usability for the purposes of this playtesting.

Following this, the lead researcher would provide the playtesters with an overview of what to expect once in the playtesting lab, describing all of the forms of data collection. In particular, the participants became familiar with the focus group questions that would be asked at the end of the session (see Section III.D.4), so that they could think about their responses to these questions during their playtesting session, yet not feel anxious about being “tested” on game knowledge following the play session.

Playtesting participants would then fill out the demographics form and sign the authorization for video/audio recording of playtesting sessions.

2) Playtesting Session: Playtesters were seated at their desks side-by-side. Each play station had a headset with earphones and a microphone, had a pad of paper and a pen, and was installed with a copy of the screen capture program. Playtesters were introduced to these features and encouraged to narrate their thought processes and/or communicate with their playtesting partners any questions or comments that arose during the session. Plenty of time was allotted for each session, to allow players time to experiment with the game and get a good feel for it. The researcher would also point out the rear-shoulder view mounted camera that was installed in the lab and instruct playtesters to wave at the camera when they were finished or if they required assistance at any point.

3) Post-playtesting Questionnaire: Following the audio/video recorded gameplay session, playtesters completed the questionnaire on the four heuristic areas. Players responded to Likert scale-type questions in these four areas, as well as filling out their comments in the corresponding section.

4) Focus Group Debrief: When both playtesters had completed the questionnaire, the researchers led a semi-structured discussion with the playtesters on what they liked about the game, what they were confused/frustrated by, and what they felt was the priority for the team to address. Researchers took detailed notes during these sessions.

IV. Implications

This start-to-finish, comprehensive method for playtesting computer games is intended to provide game developers with a malleable blueprint that can form the basis for their own playtesting protocols. Through collection of both qualitative and quantitative data and the use of player-conducted heuristic evaluation, a robust quantity of playtester data is gathered for analysis. These data can be quickly synthesized for the game developer and used to guide the development of future builds of the game.

Early and frequent playtesting data, when paired with an agile development approach, particularly as it supports the customer collaboration and responsiveness to change aspects of the Agile Manifesto [14], is of significant value to the developer creating a game that is intended for a wide audience. For serious games, the playtesting protocol described in this paper ensures that barriers to the game's teaching effects are overcome; for entertainment software and games of all genres, this approach enables the largest audience to play.

REFERENCES

[1] A. H. Jørgensen, "Marrying HCI/usability and computer games: A preliminary look," in *Proc. of the 3rd Nordic Conf. on Human-computer Interaction*, Tampere, Finland: ACM Press, 2004, pp. 393-396.

[2] T. Olsen, K. Procci, and C. Bowers, "Serious games usability testing: How to ensure proper usability, playability, and effectiveness," in *Design, user experience, and usability: Theory, methods, tools, and practice*, A. Marcus, Ed. Heidelberg, Germany: Springer-Verlag Berlin Heidelberg, 2011, pp. 625-634.

[3] T. Fullerton, "Playtesting," in *Game Design Workshop*. San Francisco: CMP Books, 2004, ch. 8, pp. 196-222.

[4] W. Barendregt, M. M. Bekker, D. G. Bouwhuis, and E. Baauw, "Identifying usability and fun problems in a computer game during first use and after some practice," *Int. Journal of Human-Computer Studies*, vol. 64, no. 9, pp. 830-846, 2006.

[5] E. Law, "Evaluation and validation methodologies for adaptive digital educational games," in *An Alien's Guide to Multi-adaptive Educational Computer Games*, M. D. Kickmeier-Rust and D. Albert, Eds. Santa Rosa: Informing Science Press, 2012, ch. 8, pp. 137-152.

[6] A. Britez. (2011, Feb. 3). What are heuristic evaluations for games? *Unthink media*. [Online]. Available: http://blog.unthinkmedia.com/2011/02/03/what-are-games-heuristic-evaluations/

[7] C. Koeffel, W. Hochleitner, J. Leitner, M. Haller, A. Geven, and M. Tscheligi, "Using heuristics to evaluate the overall user experience of video games and advanced interaction games, in *Evaluating User Experience in Games*, R. Bernhaupt, Ed. London, UK: Springer London, 2010, ch. 13, pp. 233-256.

[8] H. Desurvire, M. Caplan, and J. A. Toth, "Using heuristics to evaluate the playability of games," in *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, Vienna, Austria: ACM Press, pp. 1509-1512.

[9] J. M. C. Bastien, "Usability testing: Some current practices and research questions," *Int. Journal of Med. Info.*, to be published.

[10] J. Rubin and D. Chisnell, Handbook of usability testing: How to plan, design, and conduct effective testing. Indianapolis, IN: Wiley, 2008.

[11] J. Spool and W. Schroeder, "Testing web sites: Five users is nowhere near enough," in *Proc. of the Conf. Extended Abstracts on Human Factors in Computing Systems*, Seattle, WA: ACM Press, 2001, pp. 285-286.

[12] K. A. Ericsson and H. A. Simon, Protocol analysis: Verbal reports as data. Cambridge: Bradford Books/MIT Press, 1984.

[13] J. Miller and J. Baker-Prewitt, "Beyond 'trapping' the undesirable panelist: The use of red herrings to reduce satisficing," in *2009 CASRO Panel Quality Conference*, New Orleans, LA, pp. 1-11.

[14] Beck, Kent; et al. (2001). "Manifesto for Agile Software Development". Agile Alliance