Evaluating the Evaluation Tools

Björn Strååt
Stockholm University, Department of Computer and Systems Sciences/DSV
Forum 100
SE-164 40 Kista, Sweden
+46-709-723 222
bjor-str@dsv.su.se

Harko Verhagen
Stockholm University, Department of Computer and Systems Sciences/DSV
Forum 100
SE-164 40 Kista, Sweden
+46-8-16 16 94
verhagen@dsv.su.se

ABSTRACT

Evaluation tools specifically aimed at computer game design have been developed and adopted by the industry. Meanwhile, a recent game study by the authors of this article indicates that design choices with negative impact on player experience prevail in a number of games of both high and low rating. These results gave reason to question whether existing evaluation methods fully address all aspects of the gaming experience.

The objective of the authors is to improve existing game evaluation tools, with a focus on game experience. The new evaluations tools – i.e. heuristics – will deal with problems such as Lack of Meaningful Play (LoMP), and False Affordance (FA).

LoMP and FA are known issues in game research and Human Computer Interaction circles. These concepts need to be introduced into the area of game evaluations, since studies by the authors show that these problems are prevalent in games with low ratings.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: *valuation/methodology*; H.5.2 [**User Interfaces**] *Evaluation/methodology, User-centered design, Style guides*

General Terms

Design, Documentation, Human Factors, Standardization, Theory

Keywords

Games, Experience, Heuristics, Evaluation

1. GOALS OF RESEARCH

1.1 Introduction

Anyone who regularly plays computer games has at one point or another encountered game elements or events that have made the experience of playing the game less enjoyable. Such disruptions will impact the gaming experience to varying degrees, from passing annoyance to complete frustration. The issue of game usability problems has been thoroughly discussed by game researchers. Methods for evaluation of game design have been created and put into use by the game design industry. Federoff [1] was an early pioneer and more followed. Desurvire, Caplan and Toth [2], Desurvire and Wiberg [3] have all produced heuristic evaluation sets, which they have found that game designers have recognized and are applying on their work today.

Game designers are most likely not in the business of creating bad experiences, and they have the tools to remedy bad design. However, in a recent study (forthcoming) we found several problematic design choices in a number of games of both high and low rating.

The goal of our research is to develop new or improved heuristic values, or other tools for measuring and evaluating games and game design. One way of doing this is to examine the existing evaluation tools, and see whether and how they could be improved. This article describes our thoughts and points of view on games, design and evaluations. It also describes a study that was performed with the purpose of creating new heuristics that address game design issues that the game research community might have overlooked.

1.2 Recent Study

We performed heuristic evaluations on ten games, using heuristic values from different contemporary sources [3] [2] [4] as a base for evaluation. Only a subset of all the heuristics was used; those that we find useful in addressing issues in the game world and the interaction therein. The heuristic subset was named "The Net Heuristic List "(NHL). The purpose of the study was to see whether we could find serious issues that the existing heuristics are not addressing. We found two types of heuristics that haven't been discussed. One that covers problems with meaningful play, here called Lack of Meaningful Play (LoMP), and one that deals with False Affordance (FA).

The concept of Meaningful Play is richly described by Salen and Zimmerman [5] in Rules of Play. In a typical situation of Meaningful Play, the player has the opportunity to make a choice that will impact the game experience and the outcome of the chosen action. The importance of Meaningful Play becomes evident when a choice or action has no discernible correlation with the experience and outcome. This is an example of a situation of what we call Lack of Meaningful Play.

The concept of Affordance is described by Norman [6] as the inherent properties of an object that invite to a certain type of action with the object. In a situation where the player is led to believe that there is a possibility of interaction with objects in the game, where there is not, the player is subjected to a False Affordance.

Our findings, the method and the results of the study will be described in detail in section 3.

1.3 Heuristic Evaluations

Jakob Nielsen, one of the key founders of heuristic evaluation methods, describes [7] them as

"a method for finding the usability problems in a user interface design so that they can be attended to as part of an iterative design process. Heuristic evaluation involves having a small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics")."

A heuristic evaluation is a common method in the field of Human Computer Interaction (HCI). Nielsen's ten usability heuristics [8] are common knowledge for anyone who studies user centered design and other topics within HCI. Heuristic evaluations can be used during the design of a system, or as an evaluation method on a completed product. Heuristic evaluations should be done by more than one evaluator, where all the evaluations from the individual evaluators are aggregated into a single report. It is an informal method and should be performed during a design process or for validating a finished product. In our case, the recognized usability principles were derived from several sources [3] [2] [4]. The evaluation was performed by four individuals.

The topic of and HCI within games is strongly connected to our work. Part of the entertainment in playing games is that they are challenging and immersive - that is how they are designed. Without challenge a game quickly becomes uninteresting. Productivity software on the other hand, such as a word processor, should contain no designed challenges at all. Since most evaluation methods are developed to find (and eliminate) user challenges, it is clear to see why games need specially designed tools. Evaluating a game using generic Human Computer Interaction-evaluation tools is unfair to the game. A deliberate game challenge may be flagged as a severe error from a pure usability perspective, while it may add value to the game experience from an entertainment and game challenge perspective.

1.4 Our Intention

Our view is that there is a difference between how a game "works" and how it "feels". Evaluating a game from a strict "does it work" aspect does not answer the question what experiences it provides. Neither will it show if it is a "good" game. This gap in the existing heuristics is what we need to fill.

We want to examine games, with the existing heuristics as a tool, in an attempt to isolate design elements that makes a game less enjoyable. We want to see if the heuristics available can be developed further, and we want to know how game designers relate to heuristic evaluations and values.

1.5 Our Point of View

We claim that the expectation a player has on a game is dependent on the immersive powers of the game. The exposition of the game world contributes to immersion as it puts the player in a certain mood and awakens certain expectations.

The game world experience comes from interaction with the world, when the player meets the challenges of the world and successfully overcomes them. Challenges in the game world should be intentional, and help the player to reach the goals of the game, as well as being all part of the designers vision. Any unintentional challenge that is not part of the goal can be seen as an intrusion.

1.6 Our Approach

We suggest a natural subdivision between actual game-play (things the players do, see, expect and experience in the game world) and other things the players can or must do with the game (such as changing the graphical settings, saving, reloading, etc).

We call the actual game-play *Gameworld Interaction* (GI), and this is where the main focus of our tool development will be.

All other activities within a game we call *Supporting Interaction* (SI).

By dividing game-play from other activities a player can do, we introduce a natural division between two types of player activities where we can study the challenges of the game-play in isolation from setting up the game system.

1.7 The Net Heuristic List

Our sources [3] [2] [4] for heuristics cover many aspects in their lists. Some of their heuristics do not address GI. We decided to gather a subset of heuristics, picking the ones that suited our needs. For example "*Make effects of the Artificial Intelligence (AI) clearly visible to the player by ensuring they are consistent with the player's reasonable expectations of the AI actor.*" [3] was used while "*Player does not need to read the manual or documentation to play*" [2] was not. The Net Heuristic List contains a total of 14 heuristics from HEP [2], Play [3] and Pinelle et al [4].

1.8 Current Results

We have conducted a heuristic evaluation study on ten games. We found that all games have issues regarding GI, but lower ranked games have more issues, and more severe. Two new types of issues were also found, and we will proceed to verify these issues with game designers and players. A more complete description of this study and its results can be found in section 3.

2. BACKGROUND

2.1 Meaningful Play

2.2 Previous Heuristics and Evaluations

Several researchers have created informative lists of heuristics which have been very useful in understanding the current status of the topic.

Federoff [1] was one of the first to create a set of heuristics for games, in cooperation with a game developer team. Federoff collected heuristics from the 400 project list [9] and from literature on the subject. She compared her list with a real work place situation by following a team of game designers for a week. This resulted in a list of 40 heuristics divided into three categories; Game Play, Game Mechanics and Game Interface. This became the very first set of heuristics aimed solely at the game industry.

Sauli Laitinen [10] used Nielsen's [8] heuristics to evaluate a computer game under production. Laitinen found that the

evaluations performed with the Nielsen heuristic method were more efficient than the evaluations performed without heuristics.

However, he points out that a heuristic list tailored for games would be even more efficient.

Desurvire, Caplan and Toth [2] created the HEP (Heuristic Evaluation for Playability), where they refined the evaluation method by categorizing 43 heuristics into *Game Play*, *Game Story*, *Mechanics*, and *Usability*. The HEP is based on Federoff's work [1] and the 400 Project Rule List [9]. The HEP list carries a strong focus on game design issues and learnability of the game [11], and is, according to the authors, widely used by game designers.

Pinelle, Wong and Stach [4] identified usability issues based on game reviews on the GameSpot.com website. They grouped similar problems into categories, and created heuristics describing how to avoid the problems. Their list contains ten heuristics.

Desurvire and Wiberg [3] used a survey method on a game convention to define the PLAY heuristic list based on the HEP study. It is *"…a more refined and updated list of Game Playability Principles (PLAY)…*" The authors wanted a generalized foundation, modifiable for each game. The PLAY list contains 48 heuristics.

2.3 Challenges and Immersion

Besides previous work on heuristics and evaluation tools, we build our definition of experience and the GI on Linderoth's [12] thoughts about challenges, the ideas on immersion and experience by Jennet et al [13] and Ermi and Mäyrä's [14] categorization of immersion types .

Jonas Linderoth [12] describes how challenges in games can be expressed as either *exploratory challenges* or *performatory challenges*. Chess, for example, mainly challenges the chess player's *exploratory* skills. Most people will understand the rules on how to and be physically able to move the chess pieces on the board. Meanwhile, carefully *exploring* the options and consequences of each move is what will improve chances of winning a game of chess. On the other hand, anyone who has seen a tennis match knows that tennis is basically about striking the ball with the racquet, over the net, in a manner that impedes the opponent from striking it back. However, neither a perfect command of the racquet and ball, nor a perfect strategy, will guarantee a win in tennis. Here, the main challenge lies in the *performance*, in *performing* better than the opponent.

Linderoth uses the affordance concept (as defined Norman [6]) to describe the ability to find and execute actions - the world around us affords different actions, based on our abilities. If a performatory or exploratory challenge is not properly afforded, we will not be able to act on it and instead have a performatory or exploratory restriction, which in essence would indicate a game specific usability issue. In case a performatory challenge is difficult to understand how to carry out, it may turn into an exploratory challenge, which in turn may cause the short and valuable time to make a quick decision run out (e.g. perform the correct actions to defend from an attacking monster). This is of course true for exploratory challenges; e.g. if the player is not allowed enough time to think, the exploration may turn into a task of performance.

Jennet et al [13] describe different ways to measure immersion, and how immersion is strongly connected to absorbing game experience. They also discuss the concept of flow, and how game immersion could be considered a flow experience or not.

Ermi and Mäyrä [14] divide immersion into three categories, depending on type of experience; imaginative, sensory based, or challenge based. Imaginative immersion allows the player to use their imagination, sympathize with game characters and feel involved in the storyline. Sensory-based immersion relates to the audiovisual execution of games. Games are often aesthetically appealing and can capture the attention of even inexperienced players Challenge-based immersion is based on the interaction with the game challenges. When balanced correctly, a game challenge is not too easy to perform (which would make it dull) nor to difficult (which would make it frustrating). Challenges can be related to motor skills or mental skills, depending on game type.

3. THE STUDY

We conducted a heuristic evaluation on ten games. We wanted to test if the heuristics

see what differences we could find between games ranked high and low respectively, on www.metacritic.com. (Metacritic is a webpage which aggregates reviews from leading professional media critics. They present their data as a percentage value called a metascore). We decided that a game with a metascore of 80 or higher was a high ranked game and a game with a metascore of 40 or lower was a low ranked game.

All games were action- or role playing games released in 2013. All games were bought, installed and launched from Steam and played on PC computers of higher than recommended system requirements (Steam is a platform for digital distribution and communications developed by Valve Corporation).

Four evaluators played each of the ten games for about 3 hours per game. The exact time to stop an evaluation was decided from case to case, when the evaluator couldn't find any more issues and the game didn't introduce any new interaction styles. The evaluators were instructed to play the games as "naturally" as possible, and not use any cheats or try to "trick "the game into behaving irrationally. All issues that were found were written down into report cards, with a description on what had happened, where in the game it occurred, if it was a singular event or recurring and if it could be repeated by replaying the same section. All issues were graded for severity on a five grade scale from 0 to 4, where 0 represents a cosmetic or aesthetic issue, 2 represents annoying but playable and 4 represents such a severe issue that the player wants to quit playing. Finally, we matched the issues with the NHL list. If an issue did not match any of the heuristics on the NHL, it was set aside in a "no matching heuristic" category.

3.1 Results

The results showed that all games had issues but the severity ratings were higher in the low rank games. Several of the heuristics in the NHL list were unmatched which might indicate that these heuristics are well known and easy to remedy.

After further analyzing the issues in the "no matching heuristic"-group, we could categorize them into two new heuristics: False Affordance and Lack of Meaningful Play.

3.1.1 False Affordance

False Affordance should be understood as false information [15] – the player believes they can do something (climb a tree, pick up a

weapon) but there is no interaction possible. False Affordance was discovered in all but one game. In the high rank games it was an issue during the first stages, approximately 30 minutes, of playing. This indicates that the player needed to learn the design elements of that particular game, and once this was established, it was not a considered a problem anymore. In one of the high rank games the player wanted to pick up weapons that were lying on a table, which was not possible. Later on, when she found items that could be picked up, they were clearly indicated by a highlight and different from the non-interactive items. Once this was established, the player did not experience and more problems with false affordances.

The lower ranking games also had issues with False Affordance, but the problem was a bit more severe. For instance, when the player examined some boxes and got a highlight on one of them, which was perceived as a possibility for interaction. However, it turned out that all boxes had highlights, but not all could be interacted with. Since a lot of items, such as resources and treasure, were stored in boxes the player wanted to open as many boxes as possible, but got frustrated at the lack of clear information on which boxes to bother with.

3.1.2 Lack of Meaningful Play

Lack of Meaningful Play occurs in situations where choices and actions the player makes have no impact on the outcome, or tasks that offer no challenge have to be carried out. This issue was only present in the low ranking games, and at some instances it was so severe that the evaluators wanted to quit playing. In one game, the player is chasing a game controlled bandit on a motorbike on a highway. The player has no way of overtaking the bandit: if the player accelerates, so does the bandit, if the player decelerates, the bandit does too. The race is over after a certain amount of time, and offers no interesting challenge, and no matter what tactics the player tries to do, the outcome is the same.

We believe that these two new heuristics - False Affordance and Lack of Meaningful Play - are important. First of all because they are not part of the "old" lists, and would therefore be a possibly valuable addition. Secondly, because they were strong indicators that they were part of the overall impression of the low ranking games. If the designers of the low ranking games had designed with these two heuristics in mind, maybe their games would have scored better?

4. FUTURE WORK

We set out with the purpose of finding gaps in existing heuristics. As an interesting byproduct, we may also have gathered material that will help us encircle the common denominators for low ranking games, i.e. design elements that will impact the game experience negatively. I am convinced that going into depth in this area will enable me to produce new and even more refined heuristics and other tools for evaluation.

As a more immediate course of action, I will present our findings both to the game industry and to players. Conducting interviews with game designers will give us an idea on how our material corresponds to their thoughts and routines. We will gain insights in how they work e.g. to what extent they are using heuristics or similar methods of evaluation and what values they use to measure their design. We will learn how they relate to the issues we are presenting, and receive input and ideas for improvements of our heuristic values.

It is crucial that we also communicate our findings to players. They are the end users of any game designer's product, and we need their view: Are we overly zealous or nit picking, do we underestimate their tolerance etc. Exactly how this data collection should be done needs to be decided. Interviews might work, but a larger population needs to be heard. Desurvire and Wiberg [3] did a player survey in their Play-study, so doing a similar exercise might be a good idea.

If both players and designers agree with our findings, we are on the right track, and we can go ahead and create valid heuristics to complement the existing heuristic lists. It is also possible to create entirely new lists aimed at certain areas, or game genres.

5. REFERENCES

[1] M. A. Federoff, "Heuristics and usability guidelines for the creation and evaluation of fun in video games," Doctoral dissertation, Indiana University, 2002.

[2] 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 system*, ACM, 2004, pp. 1509-1512.

[3] H. Desurvire and C. Wiberg, "Game usability heuristics (play) for evaluating and designing better games: The next iteration," in *Online Communities and Social Computing*, Springer Berlin Heidelberg, 2009, pp. 557-566.

[4] D. Pinelle, N. Wong and T. Stach, "Heuristic Evaluation for Games: Usability Principles for Video Game Design," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008)*, 2008.

[5] K. Salen and E. Zimmerman, Rules of Play, Game Design Fundamentals, MIT Press, 2004.

[6] D. A. Norman, "Affordance, conventions, and design," *interactions*, vol. 6, no. 3, pp. 38-43, 1999.

[7] J. Nielsen, "Nielsen Norman Group," 1 January 1995. [Online]. Available: http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/. [Accessed 28 02 2014].

[8] "www.nngroup.com," 1 January 1995. [Online]. Available: http://www.nngroup.com/articles/ten-usability-heuristics/. [Accessed 01 03 2014].

[9] H. Barwood and N. Falstein, "Finitiearts," 27 02 2014. [Online]. Available: http://www.finitiearts.com/Pages/400page.html.

[10] S. Laitinen, "Do usability expert evaluation and test provide novel and useful data for game development," *Journal of usability studies*, vol. 2, no. 1, pp. 64-75, 2006.

[11] K. Isbister and N. Shaffer, Game Usability: Advancing the Player Experience, CRC Press, 2008.

[12] J. Linderoth, "Beyond the digital divide: An ecological approach on gameplay," in *Proceedings of DiGRA*, 2011.

[13] C. Jennet, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs and A. Walton, "Measuring and defining the experience of immersion in games," *International journal of human-computer studies*, vol. 66, no. 9, pp. 641-661, 2008.

[14] L. Ermi and F. Mäyrä, "Fundamental components of the

gameplay experience: Analysing immersion," in *Proceedings of DiGRA 2005 Conference: Changing Views – Worlds in Play.*, 2005.

[15] J. McGrenere and W. Ho, "Affordances: Clarifying and evolving a concept," *Graphics Interface*, vol. 2000, pp. 179-186, 2000.

[16] M. Beaudoin-Lafon, "Instrumental Interaction: An Interaction Model for Designing Post - WIMP User Interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2000, pp. 446-453.

[17] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 1990, pp. 249-256.

[18] M. Csikszentmihalyi, Finding Flow: The Psychology of Engagement with Everyday Life, Basic Books, 1997.

[19] J. Juul, A Casual Revolution: reinventing video games and their players, MIT Press, 2010.

[20] A. Rollings and E. Adams, Andrew Rollings and Ernest Adams on game design, New Riders Publications, 2003.