How Applicable is Your Evaluation Methods – Really?

Analysis and re-design of Evaluation Methods for Fun and Entertainment

Charlotte Wiberg
Department of Informatics
Umeå University
S-901 87, Umeå
SWEDEN
cwiberg@informatik.umu.se

Kalle Jegers
Department of informatics
Umeå University
S-901 87, Umeå
SWEDEN
kjegers@informatik.umu.se

Heather Desurvire
Behavioristics, Inc.
578 Washington Blvd. #179 Marina
Del Rey, CA 90292
UNITED STATES
heather3@gte.net

Abstract— **This paper presents an approach to analyze and re-design evaluation methods. The domain explored in this paper is evaluation methods for evaluating fun and entertainment. However, the approach presented may be applied in other domains as well. The approach is conceptually described and two examples of processes where the approach were used in practice are further discussed. As the map of IT applications and digital media is continuously re-designed, there is a constant need of re-designing evaluation methods.**

Keywords-usability, methods, evaluation of fun, re-design of methods

I. INTRODUCTION

Usability evaluation is an emerging methodology when it comes to entertainment and pleasure. There is a fine line separating traditional usability methods from methods measuring user experience. To address this need, there has been a development of new methods, or at least extensive use of traditional evaluations in novel ways. Earlier work has showed that empirical methods and so called inspection methods complement each other – the inspection methods are efficient when it comes to preparation, planning and analysis. However, empirical methods give the most authentic data possible. In the 1990's a number of inspection methods were developed; for instance Heuristic Evaluation (Nielsen, 1993) and Cognitive Walkthrough (Lewis et al, 1990) etc. However, the publications around the methods only include descriptions and usage of the method. There is little addressed about how they were developed, i.e. the process – the method to develop these methods. There have been attempts to mold the usability heuristics (c.f. Nielsen, 1982) towards entertainment products such as video games (Federoff (2002, 2003), Desurvire, 2004). The problem with such Heuristics is that they need to be modified, and validated for each new game that these Heuristics are utilized. In Entertainment, for video games alone, there are different modalities that will affect the Heuristics that are even applicable. These modalities that vary are for example, what type console, genre, generation of console, and so on.

In the CHI community in general there has been a need to develop a tradition of method development. In this paper we describe an approach, or conceptual framework, for how to develop heuristics to be used in expert evaluation of products and systems in two Entertainment IT domains, i.e. pervasive video games and entertainment web sites. The included studies include extensive empirical work of these industry projects. The main goal for researchers was to analyze the applicability of evaluation methods and to refine them for the next iteration. As described below, this was done as traditional evaluations of systems or artifacts, which is why a secondary goal also was present, i.e. to support the designers with the data from the evaluations of the game or web site. The process, i.e. the method to develop a method, is described in detail and the lists of heuristics are summarized briefly. Finally, a discussion about method development in general is presented.

II. CONCEPTUAL FRAMEWORK FOR DESIGNING METHODS AND HEURISTICS

Research has been conducted earlier on method development in the area of HCI (c.f. Jeffries, Miller et al. 1991; Desurvire, Kondziela et al. 1992; Karat, Campbell et al. 1992; Nielsen & Philips, 1993; Olson & Moran, 1998; Gray & Salzman, 1998; Karat et al. 1998). Therefore, it may appear that knowledge about how to analyze, evaluate, and design methods have already been developed over the years in this research discipline. However, the published research in HCI focuses mainly on the product of the research, i.e. the developed methods. Little effort is made to discuss or reflect on the process of this research. Often, the process of evaluating methods is only briefly mentioned in publications. It is seldom in focus in this type of research. Instead, the emphasis is on describing the advantages and disadvantages of the developed or re-designed methods. As no standard procedure, or conceptual framework, could be found for analysis and design of usability evaluation methods, it is important to describe how this analysis was done in the context of these studies.

The conceptual model for developing or designing evaluation methods and/or heuristic lists in the cases described in this paper can be visualized as below:
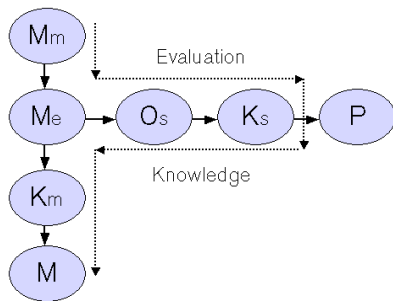

Figure 1. Two types of evaluation processes showing the evaluation method. Me= evaluation method, Os= object of study (system), Ks (knowledge about this system), P= a new (and better) product, Mm (method for studying methods), Km = knowledge about the method studied and M= a new (and better) method.

The main goal of the evaluation is to evaluate the method (Me) in itself by a method for studying method (Mm). It includes conducting a choice of an evaluation method (Me) and it is used on an object of study or system (Os). Knowledge of the system (Ks) provides us knowledge in order to build a better product (P). It also provides us with knowledge about the method (Km) which is used as input in order to design a new and more appropriate method (M) (Wiberg, 2003).

III.   DOMAIN 1- PERVASIVE GAMES

A particularly novel and promising domain of applications in the computer entertainment landscape is Pervasive Games; games that combine and make use of various user technologies (such as both mobile and stationary computing technology, cell phones, RFID technology, Embedded and Augmented reality systems, the Web, etc.) in order to create and deliver new concepts of computer gaming and new kinds of gaming experiences to players. Pervasive games differ from traditional computer games in that they heavily draw on and make use of one or several of three specific characteristics: mobile and place/time independent games (games that can be played in an anytime, anywhere fashion; at any time and at any location), social interaction as driving force in the game play (dynamic and flexible player and community created game experiences rather than pre-programmed, fixed gaming sequences and scenarios) and finally integration between physical and virtual worlds (using the physical world as part of the game environment as well as enhancing the physical world by virtual features and add-ons).

A.   SupaFly: The Pervasive Game Case

Related work on Pervasive games describes specific use of these three characteristics in game design and covers technological, software architectural and social/contextual dimensions of Pervasive gaming (e.g. Benford et al. 2005, Björk et al. 2002, Cheok et al. 2002, Flintham et al. 2003, Magerkurth et al. 2004, Manninen 2002). Since the genre of Pervasive games is in an early stage of development, knowledge considering the user/player experience in Pervasive gaming is rather limited, as well as the understanding of how the new interactions settings enabled by the Pervasiveness specifically affects User experiences.

The SupaFly case was performed as a collaborative project, conducted by researchers from Umea University, Sweden and the former game developing company It's Alive! (now part of Daydream AB), Sweden. The overall aim of the project was three-folded: to conduct methodological research considering evaluation methods for User experiences in Pervasive gaming settings, to improve the quality of the SupaFly game prototype and finally to improve the general understanding of User experiences in Pervasive gaming usage settings/situations. Further, the company wanted to improve their knowledge considering their target audiences and their intended end users, and one of the most important overall goals was to develop a method and procedure for evaluating User experiences in commercially developed Pervasive games. This latter goal includes tailoring an evaluation method and procedure to suit and meet the demands of a commercial game development process.

B.   Pervasive User Experiences- The Impetus and the Implementation

The method needed to meet the requirement of providing broad feedback from a large number of players in relatively short time (in order for the designers to get to know their target audience but also other audiences ways of reacting to the game), allow for authentic use in authentic settings, and to provide qualitative feedback; not just how much the players liked/disliked the game or how well they performed, but one step further into the "why" dimension of the gaming experiences (since both designers and users are relatively novel to the genre).

A general problem with Pervasive game prototypes is that they often are very limited in enabling the researcher to study the interaction with the environment and other users, because the use environment and corresponding Ubiquitous artifacts are not directly accessible (Manninen, 2002). Considering Usability evaluation of systems and software based on Ubiquitous/Pervasive technology, this problem generates some methodological issues. These issues became clearly illustrated in the case of the SupaFly game. First, we have problems created by the intended usage situation: since the games are intended to be played in an anytime, anywhere fashion, the possible interaction contexts includes a large number of contexts and situations in people's everyday lives. The complex and extensive number of aspects derived from the everyday life contexts makes lab studies rather problematic, especially if we would like to study the overall player-game interaction and not just the user-digital interface interaction. A possible solution to this problem would then be to use ethnographical methods and follow the users in

their everyday contexts in order to capture the player-game interaction whenever it arises. Two major problems associated with ethnography made us exclude these methods in the case: first, when the intended usage situations may occur in a very unforeseen manner we can not guarantee the integrity of the users participating in evaluation. In order to ensure the capture of all game play situations, we would need to follow the players 24 hours a day, which of course is unthinkable from an ethical perspective and not very likely to get the players' permission to do. Second, ethnographical studies was considered to be too resource consuming in order to fit the commercial development process of the game design company.

We ended up with a combination of three methods for capturing the game play and User experiences, that seemed to some extent solve the problems and fit the context of the case study: Qualitative questionnaires, System logs of user activities and Focus group interviews.

The procedure was as follows: After two weeks of free gameplay (logged as system logs) where the players experienced the game in their everyday contexts, they were instructed to fill out a qualitative questionnaire. Different players with different degrees of engagement in the game play were then invited to a Focus group interviews with selected participants. The selection was based on interesting answers (in the questionnaire) and interesting usage behavior (from system logs).

C. Results of the Evaluation- The Impact

Since the participating company, it turned out during the project, was already trying to sell their prototype to customers when the project started, the window of opportunity for redesign of the actual game was closed and therefore the possible impact of the evaluation rather limited. Besides some minor adjustments, mainly performed for business reasons and not suggested by the evaluation, the impact of the evaluation on the game was none. Some of the designers expressed that they had learned a lot about evaluation and that the results of the evaluation was very important for future projects. The main impact of the case performed was therefore of a methodological nature rather than of a practical or design process oriented; the results indicated that the procedure of applying the chosen methods turned out to be rather applicable in commercial settings given the restricted time and resources available and the type of feedback requested by the designers.

IV. Domain 2 – Entertainment services

Entertainment web sites (EWSs) are web sites that include some kind of entertainment aspects and features. They were defined in this study as to include one or more of the features listed below:

- Entertainment information – information about the theme of the web site, jokes etc.
- Downloadable items – screensavers, pictures etc.
- Small 'stand-alone' games – 'Memory or such
- Other features dependent on plug-in technology – Re-mixing of music etc.
- High quality graphic design

- Edutainent content
- Communication with others – chats, virtual meeting rooms etc.

A. The studies

These results are findings from a number of evaluations of EWSs in industry projects. Examples of web sites were Eurovision Song Contest 2000 – an event site, Mosquito – a support web site for a TV-show, The Total Defense – an edutainment web sites for teenagers about national defense etc. Overall, the group of web sites included in the studies was heterogeneous. The impetus of changes of the methods thus came from researchers. Overall, the design team strived to black box the design of- and the methods included in the studies. They were only interested in the results of the evaluations of the web site, i.e. the usability problems or implications for design regarding fun, entertainment and user experience. Overall, seven (7) entertainment web sites were included in the studies.

B. Results of the re-design - The Impact

All the studies conducted in this larger study was included in a joint collaboration project between a Swedish university and a leading Swedish web design company. The design company provided access to the web sites in their design projects and all related materials in these projects. Throughout the process the company also provided the tools and input needed to perform evaluations as part of the continuous quality assurance in its design process. The figures below originate from the design company and describe the changes they made in the design process as a result of our collaboration. The first figure shows the production process before the collaboration and the second visualizes the changed view of the same process.
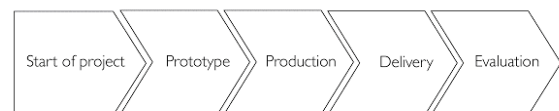

Figure 2. The production process at Paregos Mediadesign before the collaboration.
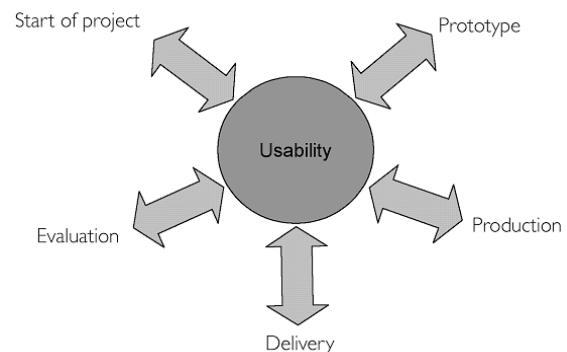

Figure 3. The production process after the collaboration.

The figures show that before the collaboration usability issues were practically not considered important in the design process. System evaluation took place during the very last phase of the production process and dealt almost exclusively with code bugs and other technological aspects. As a result of collaboration with the research team, the company changed the process, so that usability became a concern of designers at every phase of the design process. As stated by the management of the company: "Usability aspects have become a central 'measure of success' as a result of the collaboration with the research team", (the founder of the design company, 2001-10-20). The design company came to understand the importance of user-centered design, where usability is a key issue, and during the collaboration process strategies were developed to involve usability at all stages in the design process. Usability was also used from this study and on as an argument in the promotion of the company's' designs and design process.

C. *The heuristics*

Below, the final heuristics are summarized (for a complete list see Wiberg, 2003, pp. 184-185)
1. Visual impression vs. expectations
2. Exploratory design
3. Playability – Gameplay
4. Durability and lifetime – amount of content
5. Coherence between chosen design and desired mediated feeling
6. Clarity of genre – design for the right target group
7. Balance between information and entertainment
8. Originality and freshness (uniqueness)
9. Consistent navigation
10. General function related aspects

In this specific case, the design company continued to use external evaluators also after this list was developed. In that sense it did not affect the designers in their initial and/or daily work that much, i.e. they did not use it as a normative list used to make design decisions. However, evaluators used the list when conducting expert evaluations and they provided feedback to the design team which made changes when necessary. So in that sense the list did affect the designers' life.

Another result important to note in this context is the rating form that was added to the heuristic list shown above. During the project a ranking form was used continuously in order to get feedback on what heuristics were ranked as important and which were not. The main purpose here was to exclude the least ranked in every iteration of the list. What happened was that the ranking list in itself showed great potential also as a tool in evaluations when not the method was in focus. The reason is the nature of the EWSs They are very heterogeneously designed – some focus on information providing, others have a social level which is in focus and others focus on games. In order to judge the EWS, for instance, from the point of view of the playability heuristic, it seems reasonable to also judge to what extent this heuristic is important here at all. Does the web site include a game? So,

in final version, the ranking form stayed as a natural part of this method.

V. CONCLUSIONS

This paper presented a number of cases where traditional methods have been used in industry projects for two purposes, 1) to give particular feedback for designers about the design solutions, i.e. the systems, and 2) as input in the process of refining or redesigning methods. Examples of the developed methods were presented, for instance a heuristic list for entertainment web sites. It was shown how the usage of the new methods influenced design members and it was also highlighted that there is a need for a tradition in HCI to describe the development of new methods. A conceptual framework for empirically informed design of methods was presented. This potentially offers input for further investigation into how to explore, analyze and (re-)design methods in general and evaluation methods more specifically.

VI. ACKNOWLEDGEMENTS

REFERENCES

[1] Benford, S., Magerkurth, C., Ljungstrand, P. (2005) Bridging the Physical and Digital in Pervasive Gaming. In Communications of the ACM, March 2005, vol. 48. No. 3. Pages 54-57.

[2] Björk, S., Holopainen, J., Ljungstrand, P., Åkesson, K-P. (2002) Designing Ubiquitous Computer Games- A Report from a Workshop Exploring Ubiquitous Computing Entertainment. In Personal and Ubiquitous Computing (2002), 6:443-458.

[3] [Cheok, A. D., Yang, X., Ying, Z. Z., Billinghurst, M., Kato, H. (2002) Touch-Space: Mixed Reality Game Space Based on Ubiquitous, Tangible, and Social Computing. In Personal and Ubiquitous Computing (2002), 6: 430-442

[4] Desurvire, H., Kondziela, J. etal. (1992). What is Gained and Lost when Using Usability Methods Other than Empirical Testing. Proceedings of ACM CHI'1992 Conference on Human Factors in Computing Systems Posters and short talks: 125-126

[5] Desurvire, H., Capaln, M., Toth, J. (2004) Using Heuristics to Evaluate the Playability of Games. CHI 2004 Late Breaking Results Paper. In proceedings of CHI 2004, Vienna, Austria, 2004.

[6] Federoff, M. (2002) User Testing for Games: Getting Better Data Earlier. Game Developer Magazine (June 2003).

[7] Federoff, M. (2002) Heuristics and Usability Guidelines for the Creation and Evaluation of FUN in Video Games. Thesis at the University Graduate School of Indiana University.

[8] Flintham, M., Anastasi, R., Benford, S., Drozd, A., Mathrick, J., Rowland, D., Oldroyd, A., Sutton, J., Tandavanitj, N., Adams, M., Row-Farr, J. (2003) Uncle Roy All Around You: Mixing Games and Theatre on the City Streets. In the Level Up conference proceedings, Utrecht, Nederlands, 2003.

[9] Gray, W.D. & Salzman, C. (1998). Damaged Merchandise? A Review of Experiements That Compare Usability Evaluation Methods. Journal of Human-Computer Interaction. Special issue:

Experimental comparisons of usability evaluation methods. Vol. 13, No.3. 1998.(pp.203-262).

[10] [Jefferies, R., Miller, J.R., Wharton, C., Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. In proceedings of CHI'91, Conference on Human Factors and Computing Systems, New Orleans, LA. (pp.119-124.)

[11] [11] Karat, C-M., Campbell, R. et al. (1992) Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems. (pp. 397-404)

[12] Karat, J., Jeffries, R., Miller, R.M., Lund, A.M., McClelland, I., John, B:E:, Monk, A.F., Oviatt, S.L., carrrol, J.M., Mackay, W.E., Newman, W.M., Olson, G.M., Moran, T.P. (1998). Commentary on 'Damaged Merchandise?' Journal of Human-Computer Interaction. Special issue: Experimental comparisons of usability evaluation methods. Vol. 13, No.3. 1998. (pp.199-291)

[13] Lewis , C., Polson, P. et al. (1990). Testing a Walkthrough Methodology for Theory-Based Design of Walk-Up-and-Use Interfaces. Proceedings of ACM CHI'90 Conference on Human Factors in Computing Systems (pp.235-242)

[14] Magerkurth, C., Engelke, T., Memisoglu, M. (2004) Augmenting the Virtual Domain with Physical and Social Elements: Towards a Paradigm Shift in Computer Entertainment Technology. In ACM Computers in Entertainment, Vol. 2, No. 4, October 2004.

[15] Manninen, T. (2002) Contextual Virtual Interaction as Part of Ubiquitous Game Design and Development. In Personal and Ubiquitous Computing (2002), 6:390-406.

[16] Nielsen, J. (1993). Usability Engineering. Academic Press.

[17] Niesen, J. & Philips, V.L. (1993). Estimating the Relative Usability of Two Interfaces: Heuristic, Formal and Empirical Methods compared. In proceedings of INTERCHI'93, Amsterdam, The Netherlands. (pp. 214-221)

[18] Olson, G.M. & Moran, T.P. (1998). Introduction to this Special Issue on Experimental comparisons of usability evaluation methods. Journal of Human-Computer Interaction. Special issue: Experimental comparisons of usability evaluation methods. Vol. 13, No.3. 1998.

[19] Wiberg, C. (2003). A Measure of Fun. Extending the scope of web usability. Ph.D Thesis. Dept of informatics, Umeå University, Sweden.