Measuring User Responses to Interactive Stories: Towards a Standardized Assessment Tool

Ivar E. Vermeulen[1], Christian Roth[1], Peter Vorderer[2], and Christoph Klimmt[3]

[1] Center for Advanced Media Research Amsterdam (CAMeRA), VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[2] Media and Communication Studies, University of Mannheim, Rheinvorlandstrasse 5, D-68159 Mannheim, Germany
[3] Department of Communication, Johannes Gutenberg University of Mainz, Kleinmann-Weg 2, 55099 Mainz, Germany
IE.Vermeulen@fsw.vu.nl

Abstract. With the increasing number of prototypes and market applications of interactive storytelling, the understanding and optimization of how end users respond to computer-mediated interactive narratives is of growing importance. Based on a conceptual model of user experiences in interactive storytelling, a measurement instrument for empirical user-based research was developed. We report findings from an initial test of the self-report scales that was conducted with N=80 players of the adventure game "Fahrenheit". Interactivity was manipulated experimentally in order to validate the measures. Results suggest that the scales will be useful for comparing user responses to 'real' interactive storytelling systems.

Keywords: Interactive storytelling, user experience, measurement, scales.

1 The User Experience in Interactive Storytelling

While much research on interactive storytelling (IS) has been and is still dedicated to issues of technological feasibility (e.g., [1]), acceptance of future IS systems by lay audiences will depend on the satisfaction of expectations and emotional preferences. It is therefore imperative to combine psychological insight on how users respond to IS systems in order to ground design decisions and future technology developments on solid perspectives for user acceptance and market success.

Existing research on user responses to IS systems has mostly been conducted by qualitative means (e.g., [2]). While such qualitative studies have been useful in optimizing system parameters and creating more effective links between the IS world and the individual users of a given system, the measures applied do not allow acquiring standardized data for systematic testing of research hypotheses and comparing different IS systems or system versions. Quantitative measures of user responses to IS systems are thus an important yet missing tool for generating more empirical and conceptual knowledge on audience reactions and preferences.

R. Aylett et al. (Eds.): ICIDS 2010, LNCS 6432, pp. 38–43, 2010.

A standardized tool for the measurement of user responses to IS must meet two types of criteria. First, the dimensions assessed must be important from a design or theoretical point of view. Second, the measures should meet methodological quality criteria, most importantly, reliability and validity.

In order to establish the theoretical grounds for a measurement tool that meets these requirements, expert interviews and literature research in IS and media entertainment were conducted [3] [4]. As a result, a list of concepts that are proposed to play a key role in users' responses to IS systems emerged. It is organized in three main categories: A) Preconditions of meaningful user experiences, B) Common and frequent experiential qualities, and C) Concepts that reflect system-specific, individual types of responses.

The conceptual work identified five important preconditions of meaningful user experiences (part A):

- System usability, (i.e., interaction with the story must be technically smooth and error-free)
- Correspondence of system capabilities with user expectations (the system needs to convey a reasonable expectation what kind of interactive influence users can exert on the story)
- Presence (users need to establish a sense of 'being in the story world')
- Character believability (virtual agents must not damage users' illusion, e.g., through irrational behavior or poor response to user input), and
- Effectance (users must be able to recognize when and how they have causally affected the story world).

Next, a group of five types of user responses was theorized that reflect 'typical', common patterns which are likely to occur across different IS systems (part B). These frequent modes of user reactions are

- Curiosity about what will happen next,
- Suspense,
- Flow,
- Aesthetic Pleasantness (positive experiences of beauty or artistic impressiveness),
- Enjoyment (an overall sense of positively valenced experiential quality)

Finally, Part C included conceptual elements that mirror the unique characteristics of an IS system, such as the specific story content that may facilitate very diverse emotional experiences or the virtual characters that may evoke very specific user responses. Therefore, users overall emotional condition and the degree of identification with the story's protagonist were included as system-specific user reactions.

Overall, the theoretical foundation of the assessment tool therefore comprised 12 dimensions of user responses that were identified as meaningful and important across 'any' type of interactive story. Part C of the architecture allows to specify additional components that are deemed important for given systems.

2 Construction of Measures and Pilot Study Design

The conceptual ground work was translated into empirical measures by construing self-report scales for each of the 12 'standard' dimensions of user responses. Existing measures were screened and adopted or adapted as far as possible. For most dimensions, however, new scales were composed in order to achieve optimal semantic fit with the experiential dimensions and the IS system context. The assessment tool was designed for immediate administration after users' exposure to an IS system. A pilot study was conducted with these scales to examine statistical reliability (which is an important precondition for valid assessments). Moreover, a check of the scales' validity was intended that was based on an experimental manipulation of interactivity in the stimulus system.

Overall, $N = 80$ university students (22 males, 58 females; average age M = 20.08 years, SD=1.91 years) with a relatively low average degree of computer game literacy (M=1.60, SD=.84 on a scale from 1-3) participated in the study. The interactive narrative system that was used was the 2005 adventure video game "Fahrenheit" by Atari ®. While this game does not mirror contemporary technological approaches in interactive storytelling, it has been praised for its ground-breaking technology that offers an innovative level of user impact on narrative progress. As it was a ready-made product with high-quality audiovisual features, it was selected over 'true' interactive storytelling prototypes in order to avoid biased user judgments due to prototypes' underperformance in system continuity, visual appearance, or experience length.

Participants were divided into two groups. One group played the introductory sequence of "Fahrenheit" for about 30 minutes and thus actually interacted with the game and the story. The other group, however, only watched a video recording of the same game sequence that the authors had prepared in advance on the same screen. These participants thus had a non-interactive experience with the same story content. The manipulation of interactivity was applied as a potentially strong intervention that could alter user experience in ways that (some of) the instrument's scales should reflect. In Particular, the effectance scale was expected to mirror the interactivity manipulation. The recruited students were randomly assigned to the interactive or the non-interactive group. After exposure to "Fahrenheit", participants were kindly requested to fill in a computer-based questionnaire that included the 12 scales on user reactions to IS systems, as well as some demographics items. Some participants breceived credits for a course they were attending, others received 10 Euros for their participation in the experiment.

3 Results

Reliability scores of each scale were determined using the Cronbach's α coefficient for internal consistency. This coefficient indicates the degree to which the items of which one scale is to be composed actually measure the same concept in a coherent fashion. In social science research, a minimum of $\alpha = .70$ is the generally accepted convention of sufficient internal consistency (reliability). All scales met the minimal requirement, with α values ranging between .70 and .91 ($N = 80$).

The second step of analysis was an examination of how the self-report scales responded to the experimental manipulation of interactivity. Interpretable responses of specific scales were considered as initial (partial) validation of the assessment tool.

Analysis of variance (ANOVA) procedures were conducted to examine group differences between participants who had played "Fahrenheit" interactively and participants of the non-interactive condition (see table 1). Interestingly, most self-report scales did not display significant group differences. However, the effectance scale reacted to the interactivity manipulation, as people in the interactive condition reported on average higher levels of effectance than participants in the non-interactive condition. In contrast, participants in the interactive condition found the story characters to be less believable than those people who had been exposed to the non-interactive story. Likewise, participants rated the system usability significantly lower in the interactive condition than in the non-interactive condition, and they also found the experience to meet their expectations to a lesser degree.

Table 1. Results

Means and standard deviations within and significance of difference between interactive and non-interactive experiences. Note: Higher values reflect higher ratings (e.g., greater system usability), except for "correspondence with user expectations", where higher values reflect lower user satisfaction.

	Interactive Condition		Non-interactive condition		
User experiences	*M*	*SD*	*M*	*SD*	*P*
Preconditions (Part A)					
System usability (α = .84)	3.11	.94	3.69	.75	.004*
Correspondence /w user expectations (α = .81)*	3.63	.56	3.38	.62	.06
Presence(α = .91)	2.68	.98	2.62	.95	.77
Character believability (α = .76)	2.98	.90	3.48	.59	.004*
Effectance (α = .89)	3.23	.69	2.40	.97	.000*
Experiential qualities (Part B)					
Curiosity (α = .86)	3.58	.73	3.43	.64	.35
Suspense (α = .83)	3.33	.72	3.44	.77	.51
Flow (α =.74)	2.95	.71	3.00	.49	.70
Aesthetic pleasantness (α = .70)	2.00	.65	2.24	.62	.10
Enjoyment(α = .92)	2.94	.82	2.80	.66	.41
Specific experience measures (Part C)					
Emotional state: positive (α =.87)	4.60	1.66	4.51	1.50	.79
Negative (α = .90)	2.59	1.51	2.91	1.43	.33
Role adoption (α = .77)	2.71	1.04	2.67	1.05	.86

4 Discussion

With the present 12-partite set of self-report measures, a first standardized tool for the quantitative assessment of user responses to IS systems has been established based on solid theoretical ground work. The results of the pilot test with 80 players (or viewers) of the "Fahrenheit" video game suggest that the current version of the measurement tool also meets the methodological quality criteria: Internal consistency (reliability) is satisfying, for most scales rather good to excellent. Moreover, some interesting result patterns bound to the manipulation of interactivity were observed.

First, the effectance scale produced outcomes that are conceptually informative. People who were allowed to interact with the adventure game reported higher values of perceived own efficacy onto the story and the system than people who merely watched the recorded show and did not interact. This finding is of particular relevance, because effectance is conceptually very closely linked to interactivity and thus to the very core of what IS is about [5].

Next, participants in the interactive condition found characters less believable than people in the non-interactive condition. It is likely that technological limitations in character behavior produce irritations in users when they interact personally with the agent. In contrast, a video-recording of virtual characters' behavior that users only watch may let the same technical problems appear in a less irritating fashion to users. So the group difference that occurred is interpretable and suggests initial validation of the character believability scale. Similarly, the low values for system usability ratings in the interactive group seem to mirror the fact that in the interactive condition, constraints of the interaction inevitably became salient to participants. In contrast, people in the non-interactive group did not come across any usability issues at all. In that sense, also the usability scale responded in a meaningful way to the on/off-manipulation of interactivity.

And finally, the corresponding result pattern for the scale on the match between system capability and user expectations fits into this perspective as well. With the offering to participate interactively in the story events, expectations towards how the system should respond to inputs are necessarily put relatively high compared to a fully linear stimulus for which participants know that there will not be any individual interaction. Consequently, lower levels of satisfaction with what the system is capable to do are likely for the interactive condition compared to the non-interactive condition – at least as long as users have not much prior knowledge about what to expect from IS systems.

Taken together, these results suggest that the 12 subscales for the assessment of important components of the user experience in interactive storytelling meet the requirements for systematic, comparative research on IS prototypes and systems. With statistical performance values established using the "Fahrenheit" game, further studies on more advanced systems developed in the IS research community will follow to find out more about the reliability and validity of the scales. This way, scales may be optimized (e.g., by removing single items or adding subscales that are found useful completions of the overall set), and benchmarking values will be obtained for the various dimensions of user response that research teams can apply to learn more about the impact of their particular IS environment on users. Similarly, the further-tested scales will be useful for comparing different versions of a given system (e.g., different

interface types, different presentation modalities such as text versus 3D imaging) with regard to the impact of the different technological and conceptual building blocks of current and future IS systems on user experiences (e.g., [6]). Ultimately, the measurement kit that is envisioned to grow out of the present research line will then be capable to inform and guide user-oriented system development and refinement in research and application contexts.

Acknowledgment. This research was funded by the European Commission (Network of Excellence "IRIS – Integrating Research on Interactive Storytelling" – Project ID 231824). We thankfully acknowledge the Commission's support.

References

1. Cavazza, M., Lugrin, J.L., Pizzi, D., Charles, F.: Madame Bovary on the Holodeck: Immersive Interactive Storytelling. In: ACM Multimedia 2007, pp. 651–660 (2007)
2. Mehta, M., Dow, S., MacIntyre, B., Mateas, M.: Evaluating a Conversation-centered Interactive Drama. In: Conference on Autonomous Agents and Multiagent Systems (2007)
3. Roth, C., Vorderer, P., Klimmt, C.: The Motivational Appeal of Interactive Storytelling: Towards a Dimensional Model of the User Experience. In: Iurgel, I.A., Zagalo, N., Petta, P. (eds.) ICIDS 2009. LNCS, vol. 5915, pp. 38–43. Springer, Heidelberg (2009)
4. Klimmt, C., Roth, C., Vermeulen, I., Vorderer, P., Roth, F.S.: Forecasting the Experience of Future Entertainment Technology: "Interactive Storytelling" and Media Enjoyment. Presentation to the Annual Conference of the International Communication Association (ICA), Communication & Technology Division, Singapore (2010)
5. Klimmt, C., Hartmann, T., Frey, A.: Effectance and Control as Determinants of Video Game Enjoyment. CyberPsychology & Behavior 10(6), 845–847 (2007)
6. Thue, D., Bulitko, V., Spetch, M., Wasylishen, E.: Learning Player Preferences to Inform Delayed Authoring. In: AAAI Fall Symposium on Intelligent Narrative Technologies. AAAI Press, Arlington (2007)