
# D 7.2

## THE EMPIRICAL ASSESSMENT OF THE USER EXPERIENCE IN INTERACTIVE STORYTELLING: CONSTRUCT VALIDATION OF CANDIDATE EVALUATION MEASURES

| | |
|---|---|
| **Project Number** | FP7-ICT-231824 |
| **Project Title** | Integrating Research in Interactive Storytelling (NoE) |
| **Deliverable Number** | D7.2 |
| **Title of Deliverable** | The empirical assessment of the user experience in interactive storytelling: Construct validation of candidate evaluation measures |
| **Workpackage No. and Title** | WP7 - User-centered Evaluation of IS Systems |
| **Workpackage Leader** | VUA |
| **Deliverable Nature** | Report |
| **Dissemination Level** | Public |
| **Status** | Finished |
| **Contractual Delivery Date** | 30th September 2010 |
| **Actual Delivery Date** | 10 November 2010 |
| **Author(s) / Contributor(s)** | Christoph Klimmt, Christian Roth, Ivar Vermeulen, Peter Vorderer |
| **Number of Pages** | 35 |

# Table of Contents

# Abstract

Pursuing the agenda of WP7 within IRIS further, this report summarizes the construction of measurement instruments to assess the user experience in Interactive Storytelling and two experimental studies conducted to assess the instrument's reliability, validity, and practicability. Measures were developed based on the conceptual work reported earlier (D7.1.). The set of self-report scales represents a solid synthesis of technology-driven understanding of what Interactive Storytelling is about and social-scientific research on entertaining user experiences as well as standard methodology in scale construction in communication studies.

Study 1 (N = 80) applied the draft measurement tool to users of "Fahrenheit", an adventure video game with rudimentary elements of interactive storytelling and found that A) all components of the test instruments deliver satisfying results in terms of reliability and B) experimental validation results in interpretable patterns, yet more research and interpretation is necessary to fully understand the experiential processes and the validity of single measures.

Study 2 (N = 68) examined users of a widely known interactive storytelling system, "Facace" (Dow et al., 2007). Once again, the scales of the evaluation toolkit worked out well in terms of reliability. The experimental validation showed that users in the interactive condition differed from those in the non-interactive condition in terms of perceived presence, effectance, and user satisfaction, as well as in affect experienced. No differences were found for entertainment-underlying dimensions (suspense, state-curiosity, etc.), an issue that needs to be addressed in further studies.

With empirical evidence for the reliability, validity, and practicability of the measures available now from two experiments with different interactive stories, the next steps are to disseminate the ready-to-use measures to other teams of the IS research and development community and to apply the measures to other system prototypes that are being developed within the IRIS network in order to expand the range of 'benchmarking' information.

# 1. Introduction

Research and development on Interactive Storytelling (IS) is about to bring out systems and media that provide novel modes of entertainment, learning, and other experiences. The acceptance of such future IS systems by lay audiences will depend on whether they achieve the satisfaction of target audience expectations and meet user capabilities as well as emotional preferences. It is therefore important to consider 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. Moreover, social research and user responses to IS prototypes can build bridget between technology-driven research on new media systems and social science perspectives on media entertainment, learning, and other domains.

So far, existing research on user responses to IS systems has mostly been conducted by qualitative means. For instance, Mehta, Dow, Mateas, and MacIntyre (2007) confronted 12 users with the "Façade" system and collected qualitative data on specific problems that occurred in connecting user input to system reactions (e.g., how users responded if the system underperformed in understanding complex meaning in user inputs). Similarly, Aylett, Louchart, Dias, Paiva, and Vala (2005) conducted a small-scale user test with the "Fear not" system and collected children's responses with a short set of evaluation items. While such qualitative, small-scale 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 completion to existing approaches in order to generate more empirical and conceptual knowledge on audience reactions and preferences. Within the IRIS network of excellence, WP7 has been planned to develop and test such a quantitative assessment tool. Following the work plan, it is based on conceptual considerations as well as expert consultations on IS and entertainment experiences (see report D7.1. "Target Dimensions of user-centered evaluation in interactive storytelling", filed October 2009, as well as Roth, Vorderer & Klimmt, 2009).

The present report summarizes and discusses the empirical research activities conducted mainly by IRIS partner VUA to examine the first version of the empirical tool for assessing user experiences in IS. The tool is construed as a collection of 12 scales (plus an open slot for customization to specific systems) that target key requirements for meaningful user experiences (5 scales), typical manifestations of user experiences in IS systems (5 scales) and two additional elements of user responses that closely connect to the specific content, characters, etc. of the IS world under study (i.e., affective state and identification / role adoption). The 'empty slot' is foreseen for the case that system designers or evaluators intend to compliment this set of scales with an experience or response dimension that is not covered by the 12 basic components. Figure 1 gives an overview of the composition of the assessment tool.

**Figure 1: Overview of the draft dimensional architecture of the IRIS evaluation measurement toolkit (IRIS-InStET).**

| **Part A:** **Preconditions for meaningful user experiences** | **Part B:** **Common and frequent experiential qualities** | **Part C:** **Experience measures adaptable to specific systems** |
|---|---|---|
| • System usability<br>• Correspondence of system capability with user expectations<br>• Presence<br>• Character believability<br>• Effectance | • Curiosity<br>• Suspense<br>• Flow<br>• Aesthetic Pleasantness<br>• Enjoyment | • Emotional State<br>• Role adoption / identification<br>• Open Question |

After the concept-based development of the component scales (see section 2.), the WP7 research agenda foresees empirical studies that examine the values produced by the measures from users of relevant systems. The main goal of these so-called validation studies was to critically test whether the scales turn out to be statistically reliable (an important precondition for validity) and produce conceptually valid, interpretable findings. Moreover, the studies aimed to check whether the application and handling of the scales is functioning and thus practicality of the instrument is satisfying so that also other research and development teams can easily adopt the instrument for their purposes.

To achieve these research objectives, two experimental studies were conducted. The first study (section 3.) used a commercial video game ("Fahrenheit") that includes some rather simple elements of what is understood as 'real IS' in the research community nowadays, but delivers a 'full entertainment experience' in terms of complete narrative and rich audiovisual design. The second study (section 3.) used a system that is widely known in the IS community and has served as reference and source of inspiration for many teams worldwide: "Façade" (Dow et al., 2007). This 'real IS' system was chosen because it complements the game-based approach from study 1 and represents a technology that many IS research approaches can relate to.

The subsequent chapters report on these two studies, their results and our interpretation, before a general discussion and outlook on the steps ahead is offered.

# 2. Development of Self-Report Measures

From the conceptual work that had been conducted in the first phase of WP7 work together with IRIS partners and external experts on interactive storytelling (see report D7.1), measurement instruments were developed that assess the intensity or level of each of the conceptually defined elements of the user experience during exposure to an IS system. The general measurement technique of post-exposure self-report scales was selected, because A) only such a measure is capable to assess multiple dimensions without demanding too much effort for implementation – an issue particularly relevant as the final measurement toolkit is intended to support also research and development teams who are less familiar with standard social science procedures –, and B) because there are several examples of successful application of self-report measures in similar areas of entertainment research (e.g., Green & Brock, 2000; Gamelab.nl, 2010). The envisioned measurement set for the assessment of user experiences in Interactive Storytelling will thus be a questionnaire that includes scales for each of the concepts identified as relevant in the previous work phase.

The construction of items for scales followed standard procedures in communication science and psychology, which includes the search for existing measures that can be adopted and/or adapted, and the development of own items/scales for those concepts that cannot be covered by existing instruments. The VUA team thus worked through the list of concepts (see figure 1) one by one, considered existing candidate instruments for their measurement, and created new items/scales where necessary. Those new items and scales were strictly oriented to conceptual foundations from the literature in communication, psychology, and human-computer interaction (HCI). Several iterations of item discussion and improvement were conducted before the draft set of measures was considered ready for pilot testing. Because the initial studies for the examination of the scales' performance were conducted in Amsterdam, all developed and compiled items were in English language; translation into other (European) languages remained for a later work phase when initial validation would have been achieved.

Table 1 provides an overview of main sources and or conceptual foundations of the different scales included in the draft instrument used for the pilot studies. Because of the intention to tailor the measures to the specific characteristics of interactive storytelling, most instruments were modified versions of existing measures or newly developed in order to ensure semantic compatibility with the IS domain. All scales used 5-point ratings (values 1 to 5), with lower values indicating lower agreement to an item and higher values indicating greater agreement. Thus, all scales were designed to enable correlational analysis of reliability (so-called Cronbach's Alpha procedures that examine whether the component items of one scale measure the same concept) and mean value indexing for purposes of data compression (one variable reflecting one concept) and easier group comparison (e.g., for benchmarking different versions of an IS system in an experimental setting).

Overall, the IRIS user experience questionnaire (draft version) contains 96 items that measure 13 key concepts; the 'open question' for a system-specific additional component (see figure 1, section C) was left free for the purpose of initial scale testing and validation. The scales represent a synthesis of technology-driven understanding of what Interactive Storytelling is about and social-scientific research on entertaining user experiences as well as standard methodology in scale construction in communication studies.

**Table 1.Descriptions of scales employed**

| Scale | No. of Items | Example item | Main Source |
|---|---|---|---|
| *Preconditions (Part A)* | | | |
| System usability | 3 | "I thought the system was easy to use" | Adapted from Brooke (1996) |
| Correspondence /w user expectations | 11 | "I expected the experience to be more engaging" | newly developed / VUA |
| Presence | 6 | "I felt like I was part of the environment in the presentation" | Wirth et al. (2007) |
| Character believability | 4 | "I could feel what the characters in the environment were going through" | newly developed / VUA inspired by Riedl & Young (2005) |
| Effectance | 6 | "My inputs had considerable impact on the events in the game" | Klimmt et al. (2007) |
| *Experiential qualities (Part B)* | | | |
| Curiosity | 9 | "During the experience, I felt inquisitive" | Spielberger et al. (1979) |
| Suspense | 8 | "At some moments I was anxious to find out what would happen next" | newly developed / VUA based on Vorderer et al. (1996) |
| Flow | 8 | "During the experience I felt competent enough to meet the demands of the situation" | Jackson et al. (2008) |
| Aesthetic pleasantness | 5 | "I found the experience inspiring" | Adapted from Rowold (2008) and Cupchik et al. (1994) |
| Enjoyment | 13 | "The experience was gratifying" | newly developed / VUA based on Vorderer et al. (2004) |

*Specific experience measures (Part C)*

| | | | |
|---|---|---|---|
| Emotional state: positive | 10 | “At this particular moment I feel excited” | Watson et al. (1988) |
| negative | 10 | “At this particular moment I feel sad” | Watson et al. (1988) |
| Role adoption | 3 | “During the experience I felt like I was in the main character’s skin” | newly developed / VUA based on work from the FUGA project |

# 3. Study 1: Measuring User Responses to an Adventure Game with Rudimentary Interactive Storytelling

## 3.1 Context and Research Objective

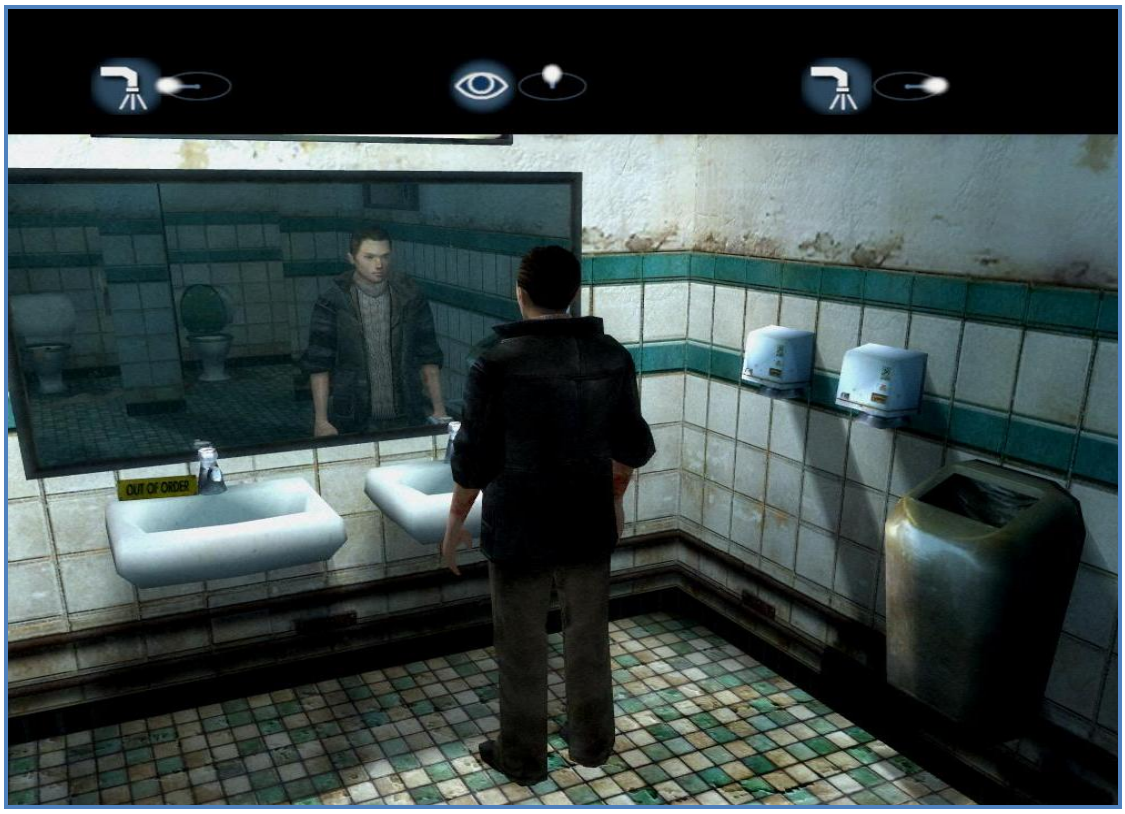
After construction of the draft version of the measurement instrument, the research agenda of WP7 was pursued further by exploring how the scales would 'perform' in pilot studies. Three issues of instrument performance were of particular interest:

- *reliability* (i.e., the extent to which the single items of each scale converge statistically to form an integrated measure of a concept that is robust against individual differences in understanding semantics and can thus be used across different studies, settings, and participant groups),
- *validity* (i.e., the extent to which the scales measure those concepts they have been designed for; this was examined through observing how the measures would respond to experimental variations of the mediated story users engaged in), and
- *practicality* (i.e., the extent to which the handling of the questionnaire in the laboratory context is functional so that researchers from different teams and different backgrounds can make quick and effective use of the instrument without investing too much time, preparatory or aftermath efforts).

For the beginning of the chain of pilot studies foreseen within the IRIS workplan, choices had to be made concerning the kind of (interactive) story used as input experience for participants and the kind of people who would be invited to serve as test audience for the measure. With regard to the interactive storytelling system to be used for the first pilot study, the various teams of the IRIS network pursue technologically very diverse approaches and are working on different interesting system prototypes. When the current pilot study was about to begin, however, none of the IRIS partners' systems appeared to be in a stage of completion (from an entertainment user perspective) that would have allowed its use in the user research context. For example, some aspects of technology that are less relevant to the IS research and development are highly relevant to entertainment users, such as advanced graphics and sound or length of experience. With the existing IS prototypes from IRIS partners (as of late 2009 / early 2010), user research would have been at risk to produce severely biased results because users' expectations towards such aspects would have been missed, whereas the 'true merits' of the interactivity of the offered narrative might have been overlooked by lay users. Related discussions about the suitability of various IRIS systems (e.g., *EmoEmma* from TEES or *idtension* from UGE) were conducted with partners and within the VUA team to make an informed decision.

Finally, no IRIS system prototype was selected for the first pilot study, but an available commercial adventure video game, "Fahrenheit" (Quantic Dream / Atari, 2005; see figure 2 for a screenshot). This games has been praised for its advanced mode of (interactive) storytelling. While its IS technology certainly does not resemble what is built-in or envisioned for the system prototypes within the IRIS network, "Fahrenheit" comes as a ready-made entertainment that is complete in the sense of graphics, sound, music, and experience length. It was thus considered to be better suited for learning about the performance of the measurement scales than an 'incomplete' IS prototype, because the primary scope of the study was not to examine interactive storytelling, but to examine the measures on user experiences under informative circumstances. Therefore, pilot studies with more advanced IS technology coming out of the labs of IRIS partners were left for later work phases within WP7.

**Figure 2: Screenshot from the adventure video game “Fahrenheit” used for study 1.**




Concerning the type of target audience to be investigated in the pilot study, the decision had to be made whether study participants should be ‘complete lay users’ with little or no experience in interactive entertainment or whether they should have some prior experience with, for instance, video games. Because the latter group is more likely a) to have an understanding of what is possible with contemporary interactive entertainment computing and what should not be expected and b) to pick up innovations in interactive storytelling early once they hit the markets, university students were recruited who were required to have at least “some” prior experience with video games. Thus, an ‘informed lay user’ audience was selected for the pilot experiment.

A final strategic decision was made concerning the mode of delivery for the measurement instrument. Because of practical advantages, the scales were not administered as a paper and pencil measure, but as a computer-based procedure. This way, the efforts of data collection and processing are reduced, error risks in data handling are minimized, and sharing the measurement instrument with other research teams becomes much easier.

## 3.2 Research Design

The examination of the measurement instrument's performance was designed as an experimental study. Participants used the "Fahrenheit" game in a laboratory setting for 30 minutes each and completed the questionnaire with the user experience scales afterwards as a computer-based procedure. The experimental component of the study was the manipulation of the story's interactivity. While half of the participants played the game in the typical way and thus interacted with the game (story) via the computer mouse, the other half of participants merely watched a pre-recorded video of the same game episode that had been created by the research team in advance. While the 'content' of the narrative remained 'constant', the interactivity of the story experiences was manipulated (on / off). It was assumed that switching interactivity of the adventure game story on or off should result in fundamental shifts of the user experience to which at least some of the instruments' scales should respond in a meaningful way. This was assumed to be the case for the effectance scale in particular, for effectance is theorized as the experiential dimension that is most directly linked to users acting in a game environment and/or story world (Klimmt et al., 2007). The experimental variation was thus implemented to produce variation in the user experiences, and the empirical question of interest was then whether and how the measurement would mirror this variation and produce meaningful patters of group differences between those participants who had engaged interactively with the story and those participants who had merely observed the story.

Overall, N = 80 university students (22 males, 58 females; average age M = 20.08 years, SD=1.91 years) with a relatively low degree of computer game literacy (M=1.60, SD=.84 on a scale from 1-3) were recruited for this experiment. Participants were randomly assigned to one of two conditions. 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 on the same screen. The video had been prepared by the research team in advance. Table 2 shows demographical measures for the participants in the interactive and non-interactive condition. After exposure to "Fahrenheit", participants were kindly requested to fill in a computer-based questionnaire that included the scales on user reactions to IS systems, as well as some demographics items. Some participants received credits for a course they were attending, others received 10 Euros for their participation in the experiment. The overall procedure typically lasted for about 50 minutes per participant.

**Table 2. Demographics of participants in study 1**

*Means and standard deviations within and significance of difference between interactive and non-interactive experiences of "Fahrenheit"*

| | Interactive condition | | Non-interactive condition | | |
|---|---|---|---|---|---|
| *Demographics* | | | | | *p* |
| Gender | 30% M<br>70% F | | 25% M<br>75% F | | .80 |
| | *M* | *SD* | *M* | *SD* | |
| Age | 20.32 | 1.93 | 19.83 | 1.89 | .25 |
| Video game literacy | 1.68 | .89 | 1.53 | .78 | .43 |

## 3.3 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 α = .70 is the generally accepted convention of sufficient internal consistency (reliability). The fourth column of Table 3 provides on overview of all scales' performance in terms of reliability. Results show that all 12 scales of the standardized assessment tool met the minimal requirement, with α values ranging between .70 and .91 (N = 80). See Appendix A for a full description of all scales employed.

**Table 3. Reliabilities of the scales of the measurement instrument of user experience in IS (study 1).**

| Scale | No. of Items | Reliability (Cronbach's α) |
|---|---|---|
| *Preconditions (Part A)* | | |
| System usability | 3 | .84 |
| Correspondence /w user expectations | 11 | .81 |
| Presence | 6 | .91. |
| Character believability | 4 | .76 |
| Effectance | 6 | .89 |
| *Experiential qualities (Part B)* | | |
| Curiosity | 9 | .86 |
| Suspense | 8 | .83 |
| Flow | 8 | .74 |
| Aesthetic pleasantness | 5 | .70 |
| Enjoyment | 13 | .92 |
| Emotional state: positive | 10 | .87 |
| Emotional state: negative | 10 | .90 |
| Role adoption | 3 | .77 |

The second step of analysis was an examination of how the self-report scales responded to the experimental manipulation of interactivity. While it was not hypothesized that all scales should reflect differences in interactivity, at least some critical elements of the measurement tool, the effectance scale in particular, were expected to be sensitive in this regard. Such a response was 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 4).

Interestingly, most self-report scales did not display significant group differences. However, as predicted, 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 ($F(1,78) = 16.7$, $p < .01$, $\eta^2 = .18$). 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 ($F(1,78) = 8.23$, $p < 01$, $\eta^2 = .10$). Likewise, participants rated the system usability significantly lower in the interactive condition than in the non-interactive condition ($F(1,78) = 8.6$, $p < 01$, $\eta^2 = .10$), and they also found the experience to meet their expectations to a lesser degree $F(1,78) = 3.4$, $p = .07$, $\eta^2 = .04$).

Finally, as far as practicality issues were concerned, the process of data collection was found to be smooth and free of handling problems. Laboratory team members did not report any technical difficulties nor any remarks by participants that would indicate comprehension problems or other issues that would put the practicability of the measurement instrument into question. The computer-based measurement procedure was thus found to operate effectively.

**Table 4. Experimental group comparisons between interactive and non-interactive engagement with the "Fahrenheit" game (study 1).**

*Means and standard deviations within and significance of difference between interactive and non-interactive experiences*

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

## 3.4 Discussion

With the present 13-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 relevant 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 that require conceptual discussion.

First, the effectance scale produced outcomes that are in line with conceptual predictions. 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 (Roth et al., 2009; Klimmt et al., 2007).

Next, the fact that participants in the interactive condition found characters less believable than people in the non-interactive condition reflects the fact that when users interact with characters, the technological limitations in character intelligence and behavior necessarily produce more irritations, interruptions, and other types of discrepancies from natural-social interaction. In contrast, a video-recording of virtual characters' behavior that users only watch 'from a distance' renders such discrepancies much less salient, because from the viewpoint of an observer, it is much easier to make sense of characters' statements and actions so that irritations are less likely to occur. In this sense, the character believability scale does not necessarily produce predicted group differences, but there is a sound conceptual interpretation to the group difference that occurred in the study.

Third, the relatively low values for system usability ratings in the interactive group can be interpreted in a similar fashion: Because there actually was an opportunity in the interactive condition to 'use' the system by entering commands, limitations in usability inevitably became salient to participants. In contrast, people in the comparison group who did not have the opportunity to interact 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 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 – regardless of how 'smart' and well-performing the interactive system actually might be.

Taken together, these results suggest that the 13 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. Most importantly, statistical reliability has been demonstrated in study 1; concerning validity, the continuation of the research line in WP7 that includes further experimental approaches will help to solidify the initial validation achieved with the present results. Further pilot tests may indicate how to optimize the scales (e.g., by removing single items or adding subscales that are found useful completions of the overall set). Together with additional efforts on validation, subsequent studies can now also deliver benchmarking values for the various dimensions of user response that other research teams within and beyond the IRIS network can apply to learn more about the impact of their particular IS environment on users.

# 4. Study 2: Measuring User Responses to a Reference System of Interactive Storytelling: “Façade”

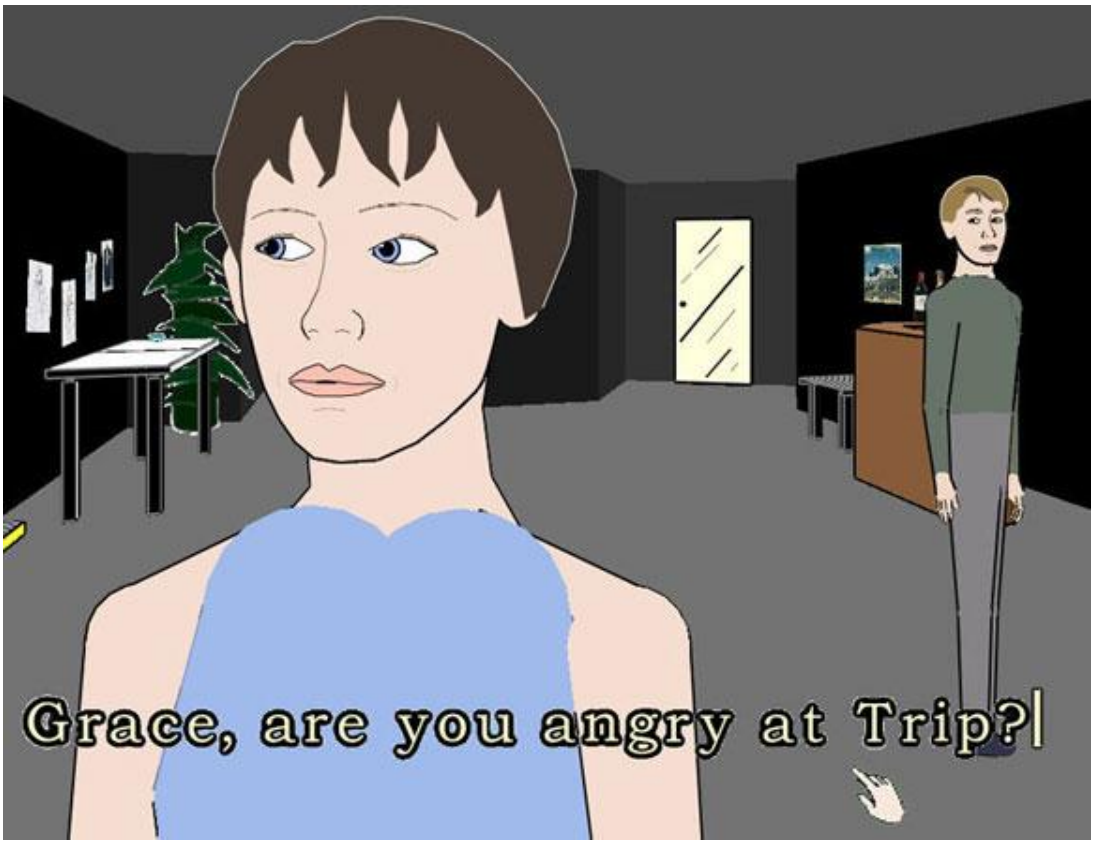
## 4.1 Context and Research Objective

The goal of the second study was to further test and optimize the developed subscales in terms of reliability, validity, and practicality. System usability had been a problem for many of our lay user participants in study 1, which may possibly explain the lack of differences observed between user experiences in the interactive and non-interactive conditions. As a result, study 1 was not able to fully answer our validity objective: Did the lack of meaningful user experiences in users of the interactive story emerge from a lack of usability, or were our measures not able to capture meaningful experiences? Moreover, the adventure video game “Fahrenheit” used in study 1 does not represent a typical case for contemporary approaches in Interactive Storytelling, because the possibilities for users to affect story development are still relatively limited. To achieve a full picture of measurement validity, additional test scores from a more advanced media environment are required that better resemble the philosophy and vision of current Interactive Storytelling.

To cope with these remaining challenges, two strategies may be employed. Concerning the usability issue, one option would be to examine, instead of relative lay users, experienced users of interactive entertainment. However, such a choice would severely limit the external validity of a study; interactive storytelling environments aim at a general audience, and not at a specific, highly experienced, audience. Representativeness of results solely obtained from an audience like the latter would be doubtful, at the least. A second strategy would be to use, as stimulus material, an interactive storytelling environment that poses less usability problems to a general audience. Testing our scales in a more basic environment would allow us to get to the core of our validation approach, namely the ability of our measures to capture the different user experiences emerging from exposure to interactively unfolding vs. pre-scripted stories. At the same time, the latter approach (using a simple yet technologically advance IS system instead of focusing on expert users) also meet the requirement of validating the user experience measures with applications more typical for modern IS technology. Because validation is key goal in our approach, we chose to pursue the second strategy and employ a highly usable interactive story environment in our second study.

In spring 2010, a review of the IS prototypes that IRIS partners were developing / working on was conducted. Conversations with team heads on IDTension (UNIGE) and EmoEmma (TEES) resulted in the assessment that these systems did not yet offer the required high levels of audiovisual appeal to lay users and also still lacked some usability for experimental application. Thus, the team VUA decided to postpone studies on original IRIS systems for interactive storytelling to a later stage of the research agenda and to use *Façade* (Mehta et al., 2005; Dow et al., 2007) for the second study. This system maintains high visibility within the IS research and development community and serves as reference system and source of inspiration for many more recent development activities. Moreover, “Facade” has been – as one of the first sophisticated IS systems ever – made available to a wider audience and was found to be highly usable and stable. Precisely because of its basic user interface, it is easy to use for relatively inexperienced users. In addition, other than the video game “Fahrenheit” that was employed in the first study, the interactive story that unfolds during play is the key feature of the game; user experiences are far less likely to be guided by impressive graphical features or overly absorbing stage settings. See figure 3 for a screenshot of Façade.

**Figure 3: Screenshot from the IS system "Façade" used for study 2.**




In sum, for study 2 the choice was made to employ a highly accessible IS system with advanced levels of IS technology to test experiences of relatively inexperienced users. Again, measurements were administered computer-based, as this approach turned out to be successful in study 1.

## 4.2 Research Design

Study 2 once again employed an experimental design Participants were invited to a laboratory at VU University Amsterdam and played Façade for about 30 minutes before completing an online questionnaire on user experiences. Similar to study 1, the story's interactivity was manipulated. Half of the participants played Façade by interacting with the story through their computer mouse and keyboard. The other half of participants instead watched a pre-recorded video of the same sequence created in advance. So, while both groups experienced the same setting and content (i.e., an argument evolving between main characters Grace and Trip), only half of the participants were able to influence the story and take matters into their own hands.

It was assumed that these different degrees of interactivity should influence the users' experiences. Following the results from study 1 as well as previous research (Klimmt et al.,

2007), users' perceived effectance was assumed to be higher in the interactive condition, possibly leading to more meaningful user experiences such as perceived presence, user satisfaction, enjoyment, and experienced affect.

In total, *N* = 68 university students (22 males, 44 females; average age M = 20.74 years, SD=5.33 years) with a relatively low degree of computer game literacy (M=1.54, SD=.74 on a scale from 1-3) participated in the experiment. They were randomly assigned to either the interactive (normal play) condition, or to the non-interactive (pre-recorded sequence) condition. Table 2 shows demographical measures for the participants in the interactive and non-interactive condition. After 30 minutes of exposure to Façade, participants filled out an online questionnaire including the scales on user reactions to IS systems, as well as some demographics items. Some participants received credits for a course they were attending, others received 10 Euros for their participation in the experiment. Similar to study 1, the overall procedure typically lasted for about 50 minutes per participant.

**Table 5. Demographics of participants in study 2**

*Means and standard deviations within and significance of difference between interactive and non-interactive experiences of "Façade"*

| | Interactive condition | | Non-interactive condition | | |
|---|---|---|---|---|---|
| *Demographics* | | | | | *p* |
| Gender | 32% M<br>68% F | | 32% M<br>68% F | | 1.00 |
| | *M* | *SD* | *M* | *SD* | |
| Age | 20.88 | 6.08 | 20.59 | 4.53 | .82 |
| Video game literacy | 1.53 | .66 | 1.56 | .82 | .87 |

## 4.3 Results

Prior to the actual data analysis, inspection of data (outlier analysis) revealed one participant scoring consistently two standard deviations below the mean a large number of items. This participant was excluded from further data processing.

The first important dimension of analysis was just like in study 1 reliability. Again, following social science conventions, Cronbach's *α* served as indicator of scale reliability, with *α* = .70 as the benchmark for sufficient scale performance. Analysis of reliability was conducted using the data from 67 participants. The fourth column of Table 6 provides an overview of all scales' performance in terms of reliability. Results show that 11 scales of the standardized assessment tool met the minimal requirement, with α values ranging between .73 and .91. Somewhat surprisingly, the widely used scale for Flow (Jackson et al. 2008), did not fully meet the requirement, at *α* = .65. Two out of the four items measuring character believability had to be deleted from the scale because of their lack of consistency with the other items. Consistency of the remaining two items was tested using Spearman's *r* correlation coefficient, which was moderate and thus acceptable. See Appendix A for a full description of all scales employed.

**Table 6. Reliabilities of the scales of the measurement instrument of user experience in IS (study 2).**

| Scale | No. of Items | Reliability (Cronbach's α) |
|---|---|---|
| *Preconditions (Part A)* | | |
| System usability | 3 | .76 |
| Correspondence /w user expectations | 11 | .84 |
| Presence | 6 | .91 |
| Character believability | 2 | .39** (*r*) |
| Effectance | 6 | .89 |
| *Experiential qualities (Part B)* | | |
| Curiosity | 9 | .84 |
| Suspense | 8 | .77 |
| Flow | 8 | .65 |
| Aesthetic pleasantness | 5 | .77 |
| Enjoyment | 13 | .89 |
| Emotional state: positive | 10 | .87 |
| Emotional state: negative | 10 | .87 |
| Role adoption | 3 | .73 |

When the reliability results from study 2 are compared to those from study 1, the similarity is striking. In fact, some reliability scores are exactly equal to the scores obtained from the first study, whereas most others are highly similar. The main exception is character believability, from which two items had to be deleted. Therefore, the scales overall turned out to be reliable once again and displayed a satisfying, mostly excellent stability across two different groups of participants who had been confronted with substantially different (interactive) stories. With some minor problems concerning the character believability and the flow scale (see chapter 5), the examination of scale reliability revealed a very good scientific performance of the measurement instrument.

Subsequently, we examined how the self-report scales responded to the experimental manipulation of interactivity. Analysis of variance (ANOVA) procedures were conducted to examine group differences between participants who had played “Facade” interactively and participants of the non-interactive condition (see table 7). Similar to study 1, some self-report scales did not display significant group differences. In the case of system usability, this was the hypothesized result: The stimulus material had been chosen to minimize usability differences between the interactive and non-interactive condition. Indeed, results showed that this difference did not occur ($F(1,66) = 0.40$, $p < .53$, $\eta^2 = .01$). Also as hypothesized, *effectance* reacted to the interactivity manipulation. Participants in the interactive condition reported significantly higher levels of effectance than participants in the non-interactive condition ($F(1,66) = 11.40$, $p < .01$, $\eta^2 = .15$). In addition, participants in the interactive condition experienced significantly higher degrees of *presence* $F(1,66) = 4.72$, $p < .05$, $\eta^2 = .07$) and *satisfaction* $F(1,66) = 5.28$, $p < .05$, $\eta^2 = .08$) than those in the non-interactive story condition. Also, the interactive condition yielded higher degrees of *enjoyment* than the non-interactive condition ($F(1,66) = 3.35$, $p < .08$, $\eta^2 = .05$) Finally, participants in the interactive condition experienced significantly more *positive affect* ($F(1,66) = 4.71$, $p < .05$, $\eta^2 = .07$) and less *negative affect* ($F(1,66) = 6.94$, $p < .05$, $\eta^2 = .10$). Table 7 provides an overview of the descriptive group comparisons.

**Table 7. Experimental group comparisons between interactive and non-interactive engagement with "Façade" (study 2).**

*Means and standard deviations within and significance of difference between interactive and non-interactive experiences*

| | Interactive Condition | | Non-interactive condition | | |
|---|---|---|---|---|---|
| User experiences | *M* | *SD* | *M* | *SD* | *P* |
| *Preconditions (Part A)* | | | | | |
| System usability | 3.93 | .81 | 3.81 | .68 | .53 |
| Correspondence /w user expectations | 3.46 | .61 | 3.10 | .66 | .025* |
| Presence | 3.27 | .84 | 2.77 | 1.00 | .033* |
| Character believability | 3.84 | .63 | 3.64 | .93 | .32 |
| Effectance | 3.18 | .92 | 2.47 | .80 | .001* |
| *Experiential qualities (Part B)* | | | | | |
| Curiosity | 3.49 | .62 | 3.33 | .78 | .33 |
| Suspense | 3.50 | .68 | 3.33 | .71 | .32 |
| Flow | 3.00 | .59 | 2.98 | .61 | .89 |
| Aesthetic pleasantness | 2.45 | .80 | 2.54 | .78 | .67 |
| Enjoyment | 2.86 | .73 | 2.54 | .73 | .07† |
| *Specific experience measures (Part C)* | | | | | |
| Emotional state: positive | 5.07 | 1.31 | 4.31 | 1.53 | .034* |
| negative | 3.05 | 1.29 | 4.06 | 1.79 | .011* |
| Role adoption | 3.24 | .80 | 2.88 | 1.02 | .11 |

Note: * significant difference at $p<.05$, † marginally significant difference at $p<.1$

## 4.4 Discussion

The results from the second study generally confirm that the set of 13 self-report measures used has satisfactory metric properties: A test with 68 users of the interactive storytelling environment Façade shows that for 11 out of the 13 measures internal consistency is satisfactory. Surprisingly, the widely used (short) scale for flow showed relatively low reliability ($\alpha$ = .65). Although reliabilities of over .60 are sometimes considered acceptable in social science research, we specifically set the acceptance level at .70, which the flow measure failed to reach. Further studies will show whether this particular result is coincidental – i.e., resulting from natural variation often occurring in social scientific data – or is more structural, e.g., denoting that perhaps the notion of flow as currently operationalized does not apply fully to interactive storytelling environments. However, given that the same flow scale met the .70 threshold in study 1 (see table 3), the interpretation of a coincidental reliability weakness seems to be more plausible. Taken the results from both studies together, also the flow scale performs at a satisfying level so far. In addition, the results suggest that the measure of character believability needs further work; Two out of the four items showed hardly any consistency with the other items. Again this pattern had not emerged in study 1 where the scale performed sufficiently reliable. Because of the importance of characters for many interactive storytelling environments and because character believability is theorized as important precondition for meaningful user experiences in IS, further examination (possibly with special experimental focus on character behaviour) is indicated. For the remaining 11 measures applied, there was a strong similarity in reliability scores to those obtained in the first study, so overall, the instrument's reliability turned out to be satisfying (and for many parts, substantially better than only satisfying) as well as stable across divergent study set-ups.

Aside from these satisfactory findings on scale reliability, the results showed that interactive and non-interactive exposures to a storytelling environment are indeed perceived as different. This conclusion could not be wholeheartedly drawn from the first experiment, because the interactive condition of "Fahrenheit" posed usability problems to our relative inexperienced audience. Our choice for a more basic interactive storytelling environment for the second study, motivated by these usability problems, paid off: The interactive use of Façade was not perceived as less user friendly than merely watching a pre-recorded sequence of the same environment.

Usability problems out of the way, the Façade study allowed us to analyze where interactive and non-interactive exposure to storytelling environments really differ in terms of user experiences. As it turns out, the basic pre-condition that perceived user effectance should be higher in the interactive condition was met: Interactive users felt significantly more influential of the development of the story. In addition, interactive users felt the environment coincided more with the expectations they held beforehand than the non-interactive users. Both these results converge with the results from study 1.

New results in study 2 were that interactive users of Façade felt significantly higher Presence in the environment than non-interactive users, and experienced more positive and less negative effect. These results can easily be explained from the finding that interactive users in the Façade study, in contrast to those in the "Fahrenheit" study, did not experience usability problems. A lack of usability will hamper feelings of immersion and absorption with a digital environment severely, and will in addition induce negative affect.

Consistent to the "Fahrenheit" study, no differences between interactive and non-interactive users were found on the dimensions underlying user entertainment, except for a marginal difference observed for enjoyment. Curiosity, suspense, flow, and aesthetic pleasantness were similar in the interactive and non-interactive conditions. Surely, this calls for further consideration. With the possible exception of flow, all measures related to enjoyment had

good metric properties, suggesting that the lack of observed difference cannot be attributed to methodological issues. In addition, the employment of the more basic Façade environment circumvented usability problems, which means that the lack of observed differences cannot be attributed to usability issues as well. Therefore, one way or another, we have to conclude that the participants in interactive condition did not perceive the environment to provide a different entertainment experience than the participants who watched a pre-recorded sequence. From a theoretical perspective of entertainment research, this observation can be explained by the fact that experiential qualities such as suspense have been argued to occur both in interactive and non-interactive settings (e.g., Klimmt et al., 2009; Zillmann, 1996). While the pathways towards an experience (such as suspense) may differ qualitatively, the actual experience measurable by our instrument might be quantitatively similar. Thus, the non-difference between interactive and non-interactive use (which partially also had occurred in study 1) does not necessarily speak against the scales' validity. Instead, it suggests that adding interactivity to a storyline does not shift all relevant experiential dimensions away from what usually happens with non-interactive stories.

Alternative explanations may also play a role here. The similarity of entertainment-related user experiences between non-interactive and interactive conditions may be due to the relative inexperience of the participants. User inexperience – e.g., game illiteracy – does not only result in problems of usability, but may also result in a failure to appreciate all the challenges and options provided by interactive story environments. Technical interactivity (a property of the IS system) does not necessarily relate into perceived interactivity and actually realized or executed interactivity. Trained users who are better able to exploit their options to affect the storyline might thus have differing entertainment-related experiences compared to the present participants. To overcome this issue, future experiments might be conducted using more experienced players, or might expose users to an interactive story environment to a much longer period of time so that the interactive features of the system have a greater chance to be perceived and actually executed by all participants.

Aside of explanations based on entertainment theory and the notion of interactivity, a third pathway of interpretation is that interactive users did in fact appreciate the challenges and options provided, but simply did not experience them as overly enticing, for example, suspense or curiosity. Possibly, interactive storytelling environments do not appeal to all audiences alike; some users may crave for the opportunity to shape a story to their own interests and ideas, while others may prefer to sit back and watch a pre-authored story unfold. Future studies should thus consider individual differences between users and include user personality measures (e.g., need for cognition, openness to experience, or need for closure) which may help explain different responses to interactive storytelling environments in different participants. Previous research on interactive movies (Vorderer, Knobloch & Schramm, 2001) supports the assumption that individual differences affect the use of and satisfaction with technical and/or story interactivity.

In sum, the findings produced by the comparison of scale values in the interactive and the non-interactive condition are conceptually meaningful and interpretable. The strong group difference in effectance (which is stable across studies 1 and 2) and the expected results on usability are particularly important cornerstones of the conclusion that also the validity of the measurement instruments has been achieved to large extents. However, open questions remain concerning the utility and configuration of the flow scale and the character believability scale.

From a technical point of view, the second experiment confirmed that ten out of 12 scales have sufficient metric properties. In addition, practicality experiences were just like in study 1 very good. From an applicability perspective, the instrument is in a very good condition already.

# 5. Conclusions and Outlook

## 5.1 General Discussion: Status of the IRIS Evaluation Toolkit

The two experiments conducted in the second phase of the WP7 agenda served successfully to mark the transition from theory work on the user experience in Interactive Storytelling to empirical measurement. The scales developed and compiled for the purpose of assessing the user experience in interactive stories were tested with two different media environments. These media applications ("Fahrenheit" and "Facade") represent a variety of those characteristics that are considered important for contemporary interactive storytelling, such as strong audiovisual immersion ("Fahrenheit") and complex, dynamic plot evolution with believable characters ("Facade"). Scientific criteria for good measures have been obtained and were found mostly good to very good in both experiments (see 3.4. and 4.4.).

The specific value of a two-study validation programme lies in the opportunity to compare how scales responded in different settings. Both studies included a manipulation of interactivity (on / off), which provides an important common ground on which scale data can be examined comparatively. The first important insight across the two studies is *stability*. The measures (with two exceptions: the flow scale and the character believability scale underperformed in study 2) were found to be reliable in both studies, which is evidence for their robustness. The demonstration of reliability in two studies is sufficient to conclude that the scales will also produce reliable results in future studies with still different (IS) settings. Stability also referred to a systematic, interpretable pattern of how the scales responded to the manipulation of interactivity across both studies. Specifically, the fact that the effectance scale – that part of the measure which most closely connects to users' ability to interact with the media environment and/or story – produced profound group differences between interactive users and non-interactive viewers in both experiments is a particularly strong evidence for the validity of the scale. Because group differences between interactive and non-interactive users emerged on several other dimensions in both studies that were conceptually reasonable and interpretable in a meaningful way, the critical question whether the scales operate effectively in an interactive context can be answered with "yes". So validity was also obtained as a stable, cross-experiment pattern, which clearly indicates that instrument development is on a good way.

In addition to the stability aspect, comparing the scale data from the two experiments also allows to reflect on some informative differences between the studies. First, "Facade" was rated as substantially higher in usability in the interactive conditions (M = 3.93 versus 3.11 in "Fahrenheit"), which mirrors important design differences between the stories, as "Fahrenheit" as an interactive movie or adventure game comes with much more complicated affordances for users when to interact and how to interact in given game/story situations than "Facade". The fact that the usability scale "detected" the fundamental differences in system usability is thus an important further aspect of overall scale validity. Second, character believability ratings were much higher in interactive "Facade" users (study 2, M = 3.84) than in interactive "Fahrenheit" users (study 1, M = 2.98). This difference can be explained by the fact that the interaction among characters in "Facade" is much more focused on a limited plot (a relationship argument between the protagonists Trip and Grace), and "Facade" characters were designed with much effort to act intelligently and believably. In contrast, "Fahrenheit" contains a crime drama story with exceptional events (such as the protagonist finding a dead body under his hands) and actions (e.g., escaping the police), which lets appear the "Fahrenheit" protagonist much less "life-like". Again, this difference between the experiments supports the assumption of scale validity. Third, "Presence" was also higher among interactive "Facade" users (study 2, M = 3.27) than among interactive "Fahrenheit" players

(study 1, M = 2.68); interestingly, this finding occurred although “Fahrenheit” comes with more elaborate graphics, sound, and dynamics. However, “Facade” puts the user in the midst of a dense interpersonal conflict, and the intelligent design of “Facade” affords higher levels of (social) Presence, which the scale data reflect. Again, this interpretable finding provides support for the claim of validity for the measurement instrument. Finally, it is noteworthy that “Facade” also created more negative affect (M = 3.05 in the interactive condition) than “Fahrenheit” (M = 2.59, study 1, interactive condition). Given the fact that “Facade” is about a topic that is not funny (relationship argument), this difference again nicely reflects design differences of the two interactive stories examined in the two studies. Overall, then, not only the stability aspect, but also the aspect of observed differences across experiments allows interpreting the measurement instrument as valid and useful for the examination and ‘benchmarking’ of interactive story prototypes and systems. In fact, it will be highly interesting to compare scale data obtained from users of other interactive stories with the reported findings from studies 1 and 2.

Another important commonality of the two studies is the very good practicality of the instrument. In both experiments, application went smoothly, and comprehension errors did not occur. Notably, we used an English questionnaire with Dutch student samples for both studies. The English version can thus also be used for other populations that do not speak English as native language (but have still good English capabilities, of course). The strategy to desight the measures as computer-based (online) assessment has contributed to the good practicality, and it will now be very useful for dissemination of the scales and their application in remote contexts (see 5.2.)

In terms of other findings of the reported two-study programme, an important finding that is interesting beyond issues of social-scientific instrument quality is that several enjoyment-related experiential components (such as curiosity, surprise, suspense) as well as enjoyment itself did not display substantial differences in interactive and non-interactive users. While this does not imply that these dimensions are irrelevant to the IS experience, it calls for theoretical explanations (see 4.4) that may also have an effect on how to envision user experiences with full-scale interactive stories of the future. Maybe some experiential qualities of IS use do not depart from what is well-known in fiction readers (e.g., Oatley, 1994) or movie viewers (e.g., Oliver & Bartsch, 2010). If suspense, curiosity, flow, and other elements of the entertainment experience delivered through IS systems do not differ so much (quantitatively) from conventional media, this may afford alternative design strategies and/or new ways of thinking what to make users expect from IS, for instance, in user instruction before the use of an IS prototype, or in marketing of full-scale IS systems to mass audiences in the future. Alternatively, more qualitative work could shed some light on whether phenomenological differences between interactive stories and conventional stories are ‘hidden’ behind that numeric data obtained by the current instrument. However, exploratory research (that used the think aloud-method: Klimmt, Vorderer & Nuss, 2010) as well as experimental research – both with video gamers - suggest that the experiential categories users apply to the interactive narrative experience do mirror conventional categories quite directly. So the invariability of enjoyment and its close relatives among the experiential components under investigation requires further theoretical and empirical consideration. With the present data, an exploration of which experiential components actually predict enjoyment of IS (in the interactive and in the non-interactive case) could serve as promising next step of data analysis.

Overall, the reported experiments with 148 participants have been successful in determining solid quality of the developed measurement instrument. They have also shown the applicability of the scales across different media environments and experimental settings. And they have in addition produced conceptually interesting results concerning how to model the user experience in IS. Compared against the work plan of IRIS WP7, the second phase of research on the user experience has thus been completed successfully – albeit some minor issues need to be resolved with respect to the flow scale and the character believability scale.

Consequently, the next steps of the WP7 agenda can be pursued with ‘positive energy’ from the successful testing and validation of the measurement instrument.

## 5.2 Outlook: Next Steps

With an established measurement instrument plus ‘benchmarking data’ available, the remaining work of WP7 will pursue three objectives:

- Exploiting the chances of participating in the IRIS network to a greater extent concerning empirical research on IS user experiences,
- Testing other IS prototypes of enlarge the range of empirical results and benchmarking data, and
- Compiling a ‘toolkit box’ of the measures that enables other research teams in the IS community to apply the measures without great effort, high level of robustness and convenience.

Further experiments with interactive stories of other kinds than “Fahrenheit” and “Facade” are foreseen to address the first two objectives. Currently, a joint experiment with UoA is being planned that will apply the scales in the context of a newly developed IS prototype. Further studies with media applications or prototypes from IRIS partners are envisioned for 2011. The third phase of WP7 work may also include another experiment on a most recent video game with advanced interactive storytelling features such as “Heavy Rain”. The rationale behind the next studies will be to obtain a larger database of results produced with the IRIS user experience measures that can be documented and serve other research teams for for comparison purposes. Moreover, a synthesis of all study results (to be conducted towards the end of the IRIS working period) will also allow to draw more profound conceptual conclusions on theorizing user experiences in IS and building new bridges to entertainment theory in media psychology and to research on Human-Computer-Interaction (HCI).

In addition to applying the measurement instrument in further studies, preparation of a ready-to-use toolkit will begin in early 2011. The goal is to compile the computer-based measure in a universally applicable format together with a template for data analysis (i.e., computation procedures for determining scale reliability and construction of scale mean values), a documentation of benchmarking values (including the findings from the present experiments), as well as instructional material on how to use the toolkit. This “toolbox” shall then be disseminated widely through a workshop at the ICIDS conference 2011, online communication (with research teams in Europe and beyond being addressed actively), and potentially further methods that benefit from support of other IRIS partners. For all these activities – more experimentation with IS prototypes and systems as well as compilation of the measurement toolbox – the present research provides the methodological foundation, at it has critically examined and successfully demonstrated the reliability, validity, and practicality of the IRIS measurement instrument for user experiences in Interactive Storytelling.

# 6. References


Aylett, R. S., Louchart, S., Dias, J., Paiva, A., Vala, M.: FearNot! - An Experiment in Emergent Narrative. In T. Panayiotopoulos et al. (eds.) Intelligent Virtual Agents: 5th International Working Conference (IVA 2005 Proceedings), pp. 305--316. Springer, Berlin (2005)

Brooke J.: SUS - A quick and dirty usability scale. Usability evaluation in industry, pp. 189--94 (1996)

Chupchik, G. C., Lázló, J.: The Landscape of Time in Literary Reception: Character Experience and Narrative Action. Cognition and Emotion, 8, pp. 297--312 (1994)

Dow, S., Mehta, M., Harmon, E., MacIntyre, B. & Mateas, M.: Presence and engagement in interactive drama. Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 1475--1484. New York: ACM (2007)

Jackson, S., Martin, A., Eklund, R.: Long and Short Measures of Flow: The Construct Validity of the FSS-2, DFS-2, and New Brief Counterparts. Journal of Sport and Exercise Psychology, 30, pp. 561--586 (2008)

Klimmt, C., Hartmann, T., Frey, A.: Effectance and Control as Determinants of Video Game Enjoyment. CyberPsychology & Behavior, 10 (6), pp. 845--847 (2007)

Klimmt, C., Vorderer, P. & Nuss, S.: Interactivity versus narrative: Using think-aloud data to understand the enjoyment of playing adventure video games. Panel presentation at the Annual Conference of the International Communication Association (ICA), Game Studies Interest Group, 22.-26. June 2010, Singapore.

Klimmt, C., Rizzo, A., Vorderer, P., Koch, J. & Fischer, T.: Experimental evidence for suspense as determinant of video game enjoyment. Cyberpsychology and Behavior, 12 (1), pp. 29--31 (2009)

Mehta, M., Dow, S., MacIntyre, B., Mateas, M.: Evaluating a Conversation-centered Interactive Drama. Conference on Autonomous Agents and Multiagent Systems (2007)

Oatley, K. A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. Poetics, 23, pp. 53--74 (1994)

Oliver, M. B. & Bartsch, A.: Appreciation as audience response: Exploring entertainment gratifications beyond hedonism. Human Communication Research, 36 (1), pp. 53--81 (2010)

Riedl, M., Young, M.: An objective character believability evaluation procedure for multi-agent story generation systems. In T. Panayiotopoulos et al. (eds.) Intelligent Virtual Agents: 5th International Working Conference (IVA 2005 Proceedings), pp. 278--291. Springer, Berlin (2005)

Roth, C., Vorderer, P., Klimmt, C.: The Motivational Appeal of Interactive Storytelling: Towards a Dimensional Model of the User Experience. In: Iurgel, I., Zagalo, N., Petta, P. (eds.) International Conference on Interactive Digital Storytelling, ICIDS, Springer, Guimarães (2009)

Rowold, J.: Instrument development for esthetic perception assessment. Journal of Media Psychology: Theories, Methods, and Applications, 20 (6), pp. 35--40 (2008)

Spielberger, C., Jacobs, G., Crane, R., Russell, S.: Preliminary manual for the state-trait personality inventory (STPI). Unpublished manuscript, University of South Florida, Tampa (1979)

Vorderer P., Klimmt C., Ritterfeld U.: Enjoyment: At the heart of media entertainment. Communication Theory, 14 (4), pp. 388--408 (2004)

Vorderer, P., Knobloch, S. & Schramm, H.: Does Entertainment suffer from Interactivity? The impact of watching an interactive TV movie on viewers' experience of entertainment. Media Psychology, 3(4), pp. 343--363 (2001).

Vorderer, P., Wulff, H. J., Friedrichsen, M. (eds.): Suspense: Conceptualizations, Theoretical Analyses, and Empirical Explorations. Mahwah, NJ: Lawrence Erlbaum Associates (1996)

Watson, D., Clark, L. A., Tellegen, A.: Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. Journal of Personality and Social Psychology, 54, pp. 1063--1070 (1988)

Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., Saari, T., Laarni, J., Ravaja, N., Ribeiro Gouveia, F., Biocca, F., Sacau, A., Jäncke, L., Baumgartner, T. Jäncke, P.: A Process Model of the Formation of Spatial Presence Experiences. Media Psychology, 9 (3), pp. 493--525 (2007)

Zillmann, D.: The psychology of suspense in dramatic exposition. In P. Vorderer, H. J. Wulff & M. Friedrichsen (Eds.): Suspense: Conceptualizations, theoretical analyses, and empirical explorations (pp. 199--231). Mahwah, NJ: Lawrence Erlbaum Associates (1996)

# 7. Appendix: Scales and Items of the IRIS Evaluation Toolkit (as used in evaluation studies)

**Curiosity** (as a user experience)

During the experience …

1. …I felt like exploring my environment.
2. … I felt curious.
3. … I felt interested.
4. … I felt inquisitive.
5. … I felt eager.
6. … I felt in a questioning mood.
7. … I felt stimulated.
8. … I felt disinterested. (N)
9. … I felt mentally active.
10. … I felt bored. (N)

- adapted from STPI, Spielberger et al, 1979
- Likert scale 1-5, ranging from “strongly disagree” to “strongly agree”
- (N) denotes a negatively framed item; recoding needed before analysis

**Flow**

During the experience…

1. …I felt competent enough to meet the demands of the situation
2. …I acted spontaneously and automatically without having to think
3. …I had a strong sense of what I wanted to do
4. …I had a good idea while I was performing about how well I was doing
5. …I was completely focused on the task at hand
6. …I had a feeling of total control over what I was doing
7. …I was not concerned with how others may be evaluating me
8. …the way time passed seemed to be different from normal
9. … I found it extremely rewarding.

- FSS-2; Jackson, Martin, Eklund, 2002
- Likert scale 1-5, ranging from “strongly disagree” to “strongly agree”

**Suspense**

1. At some moments I was anxious to find out what would happen next
2. I was really hoping that the choices I made would work out well
3. I didn't care less how the story developed (N)
4. I found myself staring at the screen in anticipation
5. Sometimes I was worried about how the story would develop
6. Some moments were rather suspenseful
7. At some points I breathed a sigh of relief
8. I found myself wishing for a particular story outcome
9. The story did not affect me (N)
10. At some points I was afraid that things would go wrong

- VUA, 2010
- Likert scale 1-5, ranging from "strongly disagree" to "strongly agree"
- (N) denotes a negatively framed item; recoding needed before analysis

**Aesthetic pleasantness**

The experience…

1. …made me think
2. …made me think about my personal situation
3. …told me something about life
4. …was inspiring
5. …moved me like a piece of art

- Adapted from Rowold, 2008, and Cupchik & Laszlo, 1994
- Likert scale 1-5, ranging from "strongly disagree" to "strongly agree"

**Enjoyment**

The experience…

1. …was pleasant
2. …was gratifying
3. …was rewarding
4. …was amusing
5. …was exhilarating
6. …was thrilling
7. …was exiting
8. …was melancholy
9. …was moving
10. …was appealing
11. …was pleasing to the senses
12. …made me feel proud
13. …made me feel competent

- VUA, 2010
- Items 1-3 are general, items 4-13 capture different facets of enjoyment (Vorderer e.a., 2004):
    - 4-5 amusement, 6-7 suspense, 8-9 melancholy, 10-11 aesthetics, 12-13 achievement
- Likert scale 1-5, ranging from “strongly disagree” to “strongly agree”

**Affect**

How do you feel at this moment, after experiencing the story?

1. Interested
2. Sad (N)
3. Excited
4. Troubled (N)
5. Powerful
6. Guilty (N)
7. Scared (N)
8. Hostile (N)
9. Enthusiastic
10. Proud
11. Annoyed (N)
12. Alert
13. Ashamed (N)
14. Inspired

15. Nervous (N)
16. Determined
17. Careful (N)
18. Hysterical
19. Lively
20. Anxious (N)

- PANAS, Watson e.a., 1988
- Likert scale 1-10, no labels
- (N) denotes items measuring negative affect

**Role adoption/identification**

1. I felt like a was in the main character's skin
2. I sometimes forgot about myself because I was so focused on the actions of the main character
3. I felt more like the character than like myself

- Adapted from FUGA project, 2010
- Likert scale 1-5, ranging from "strongly disagree" to "strongly agree"

**System usability**

1. I thought the system was easy to use.
2. I would imagine that most people would learn to use this system very quickly.
3. I found the system very cumbersome to use. (N)

- Adapted from Brooke, 1996
- Likert scale 1-5, ranging from "strongly disagree" to "strongly agree"
- (N) denotes a negatively framed item; recoding needed before analysis

**Correspondence of system capability with user expectations/User satisfaction**

1. The experience was better than I expected
2. I probably expected to much from the experience (N)
3. I was satisfied with how the system performed
4. I expected the system to be more user-friendly (N)
5. I expected the experience to be more immersing (N)
6. I expected the story's characters to be more believable (N)
7. I expected to have more control over the experience (N)
8. I expected the experience to be more surprising (N)
9. I expected the experience to be more thrilling (N)
10. I expected the experience to be more engaging (N)
11. I expected the story to be better (N)
12. I expected the graphics to be better (N)
13. I expected the experience to be more enjoyable (N)

- VUA, 2010
- Items 1-3 relate to general expectations, items 4-13 relate to expectations about specific IS facets :
    - 4 usability, 5 presence, 6 character believability, 7 effectance, 8 curiosity, 9 suspense, 10 flow, 11-12 aesthetic pleasantness, 13 enjoyment
- Likert scale 1-5, ranging from "strongly disagree" to "strongly agree"
- (N) denotes a negatively framed item; recoding needed before analysis

**Character believability (VUA, 2010)**

1. I could feel what the characters in the environment were going through
2. I had the impression that the characters in the environment responded in a thoughtful way to what I did
3. I noticed when the characters in the environment displayed strong emotions
4. The characters in the environment seemed to have a strong will of their own

- VUA, 2010
- Likert scale 1-5, ranging from "strongly disagree" to "strongly agree"

**Effectance**

1. My inputs had considerable impact on the events in the story
2. I had the feeling that I could affect directly something on the screen
3. The consequences of my inputs were clearly visible
4. I could recognize which events in the story I have caused with my inputs.
5. My decisions clearly influenced how the story went on.
6. I discovered how my earlier actions influenced what happened later in the story.

- Adapted from Klimmt, e.a., 2007
- Likert scale 1-5, ranging from “strongly disagree” to “strongly agree”

**Presence**

1. I felt like I was a part of the environment in the presentation.
2. I felt like I was actually there in the environment of the presentation.
3. I felt like the objects in the presentation surrounded me.
4. It was as though my true location had shifted into the environment in the presentation.
5. I felt as though I was physically present in the environment of the presentation.
6. It seemed as though I actually took part in the action of the presentation.

- Selected items from MEC Presence questionnaire, Vorderer, 2004
- Likert scale 1-5, ranging from “strongly disagree” to “strongly agree”

**Open questions**

1. Do you have any further questions or remarks? (open, about 5 lines)
2. Did you like the story? (open, about 5 lines)

**Demographics**

1. Gender (M/F)
2. Age (years)
3. How much experience do you have with video games? (beginner, moderate, experienced)
4. Did you play this game before? (yes/no)