D 7.3

REPORT ON BENCHMARKING AND FULL-SCALE DEMONSTRATIONS OF EVALUATION TOOLKIT MEASURES

Project Number	FP7-ICT-231824
Project Title	Integrating Research in Interactive Storytelling (NoE)
Deliverable Number	D7.3
Title of Deliverable	Report on benchmarking and full-scale demonstrations of evaluation toolkit 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 June 2011
Actual Delivery Date	July 10, 2011
Author(s) / Contributor(s)	Christoph Klimmt, Christian Roth, Ivar Vermeulen, Peter Vorderer
Number of Pages	36

Table of Contents

Abstract

The main objective of IRIS WP7 is to develop a standardized methodology for user-centered evaluation research on IS prototypes and media. Based on initial theory work and expert interviews, a set of dimensions of user experiences had been assembled and translated into a questionnaire instrument (D7.1; D7.2). With these (preliminary and revised) self-report measures, a variety of studies has been conducted that assess user responses to diverse kinds of interactive digital stories. The present report summarizes the findings about how the measurement scales performed across VUA's most recent four thematic studies: Participants were confronted with "Facade", "Emo Emma" (from IRIS partner TEES), the "Virtual Beergarden" (from IRIS partner UOA) and the role playing game "Fable: Lost Chapters". Findings on the scales' quality mark the completed empirical-statistical foundation of the IRIS evaluation toolkit.

In addition, a comparative overview of average user responses across the tested different IS systems (e.g., Facade, EmoEmma, Virtual Beergarden) is presented. It is based on all six studies that involved 316 participants and five different IS systems and prototypes. An overview table provides reference data for future applications of the toolkit (both within and outside of IRIS) and allows conceptual conclusions about important and potentially less relevant aspects of user experiences in future IS media. Within WP7, these materials will serve as base for a software application that supports research teams in collecting and analyzing standardized data on user experiences in IS environments.

1. Introduction

One of the key scientific objectives of the IRIS Network is to support the research community in finding out more about user experiences and preferences concerning interactive stories. User-centered research is both of theoretical value (i.e. understanding the kind and quality of entertainment experiences that interactive stories can foster) and of practical relevance (i.e., important information to optimize design approaches and prototypes to meet user expectations and valuations). Because the conceptual base for conducting user research with interactive stories was sparse when IRIS was planned, WP7 is about developing solid, theory-grounded measures that can be applied in user studies. Instead of 'evaluating' given prototypes with a standardized protocol, the goal is to come up with (1) a conceptual understanding of which dimensions of user experiences to look at (which is informed by entertainment research in media psychology) and (2) self-report measures that translate this conceptual understanding into empirical research. The measures are envisioned to help IS researchers and creators in focusing on user aspects that are particularly relevant for their individual approach or system. So the measurement toolkit is not a list of must-do questions, but rather a set of scales assessing different experiential response from which a given research team can select according to specific research topics.

A good self-report measure needs to be tested and validated before it can serve the research community as a basis for standardized application. Thus, VUA has together with several IRIS partners carried out six experimental studies that produced data on the various scales included in the toolkit. These data provide means to establish the statistical quality of the measures (reliability, stability), experimental validation of (parts of) the scales, and a gallery of reference values with which future research results can be compared in order to determine specific strengths and weaknesses of the examined IS system.

The present report summarizes the findings from WP7 user research that involved the scales of the measurement toolbox. With different experimental designs, a broad range of different IS systems was used for these studies: "Facade" (Dow, Mehta, Harmon, MacIntyre, & Mateas, 2007; two studies), the Virtual Beergarden (from IRIS partner UOA), EmoEmma (from IRIS partner TEES), as well as narrative-rich video games "Fahrenheit" (Atari, 2005) and "Fable: Lost Chapters" (Lionhead Studios / Microsoft, 2005). We first outline the methodology and results of the latest WP7 experiments (section 2). Subsequently, a comparative overview of how the scales performed across the six studies is assembled and discussed (section 3). Finally, the status of the IRIS evaluation toolkit for user-centered research is reviewed, and the next steps towards the final toolkit software solution are outlined (section 4). The appendix offers the latest version of the toolkit scales in original English and German wordings.

2. The Final Round of Experimental Studies in WP7

During the time period covered by this report, VUA has conducted a set of experiments that complete the first round of studies for development and validation of user experience measures. So far, findings from an experiment that employed the adventure video game "Fahrenheit" and from another experiment with the foundational (in a sense 'historic') interactive storytelling system "Facade" had been documented. These studies had revealed that virtually all elements of the measurement toolkit work well in a statistical sense (reliability, stability). Subsequent to these studies, the scope of further empirical work was broadened in two ways: More different interactive storytelling systems were involved, and more specific research issues were addressed. Overall, four additional studies have been conducted, which are briefly summarized in this chapter: One study again employed "Facade", but this time inspected test/retest-stability of the measures by having participants use the system and respond to the questionnaire twice (see section 2.1.). Another experiment was conducted around a prototype from IRIS partner UOA, the "virtual beergarden", and compared two different modes of dialogue-based user interaction with the system (see section 2.2.). The stimulus of the third reported experiment was a role playing video game ("Fable: Lost chapters") that was particularly suited to study issues of autonomy and degrees of freedom in user decisions on story progress (section 2.3.). And finally, VUA collaborated with TEES on applying the measurement toolkit to the "EmoEmma" system (section 2.4.). From this set of studies and the previous experiments reported earlier, a treasure of empirical data is available that allows drawing conclusions on the methodological strengths and weaknesses of the measurement toolkit. These conclusions are discussed in the final section of this chapter (2.5.)

2.1 A Study on Retest-Stability using "Facade"

Background and Scope

One important element in assessing self-report measures' statistical performance is to demonstrate its stability across repeat applications. Within the WP7 empirical research program, stability is inspected as a by-product of the multi-study, multi-system approach (see section 3). However, a particularly informative assessment of scale stability is a repeat measurement design in which the same participants respond to the measures more than once (so-called retest stability, cf. Kline, 1986). While in repeated measurement situations, response behaviour of a single person may vary due to situational circumstances, such as changes in mood, affect, stress, or different experimental stimuli, a substantial level of stability can be expected for many concepts to be assessed that is bound to the personality characteristics of the respondent. For example, people who tend to reach high levels of arousal very quickly should respond to two different arousing situations similarly, that is, display a similar high level of arousal in both situations, whereas a person with a lower arousability should display a lower level of arousal in both situations. Such stability must be reflected in the statistical performance of self-report measures.

Therefore, one WP7 study was dedicated to intra-person repeat measurement application. We once more selected the foundational prototype for modern interactive storytelling approaches, "Facade", because it is a well-known reference system in the field and had turned out as valuable in the previous study (see D7.2.).

Method

To compare participants' experiences following first and second exposure to the IS environment Façade, we employed a within-subjects factorial design with two conditions. A total of 50 university students (17 males, 33 females; average age M = 19.8 years, SD=1.73 years) with a low to moderate degree of computer game literacy (M=1.78, SD=.71 on a scale from 1-3) participated in the study. Three participants were excluded from data analysis because of implausible response patterns in the questionnaires.

Upon arrival in the laboratory, participants interacted with Façade for 20 minutes (first exposure). Next, they completed the IRIS toolkit questionnaire including demographical questions, and questions relating to the 13 proposed user experience dimensions in the following order: Curiosity, Suspense, Flow, Aesthetic pleasantness, Enjoyment, Affect, Role adoption, System usability, User satisfaction, Character believability, Effectance, Presence, Pride. Scales were considerably shortened to 2-item or 3-item versions in order to enable repeat application within one laboratory session. In addition, tow open questions about their experiences were asked. Subsequently, participants proceeded to interact with Façade for another 20 minutes (second exposure), after which they completed the same questionnaire once again (excluding the demographical questions). The written transcripts from both interactions with Façade were recorded and saved. Upon completion of the second toolkit questionnaire, participants were thanked, received 15 EUR as compensation, and were debriefed and dismissed.

Results

Because the scales measuring each of the 13 user experience constructs had been considerably shortened, it was particularly important to test whether these scales hold up in terms of reliability. In social science research, a minimum of $\alpha = .70$ is the generally accepted convention of sufficient reliability, although a minimum of $\alpha = .60$ is also often mentioned in the literature.

Reliability analysis show almost all scales of the standardized assessment tool met the minimal requirement following both rounds of exposure, with α values ranging between .66 and .89 (N = 50). The only exception is the reliability for the negative affect subscale following the first exposure, which was .57. Reliabilities of the 2-item scales of user satisfaction, character believability, effectance, suspense, enjoyment, and role adoption were assessed using a Pearson's correlation, and were also sufficient. For some scales, elimination of one item helped to improve reliability; the suspense scale was reduced from four to two items to generate a sufficiently reliable measure. Table 1 describes the measurement instruments used, including numbers of items (with the number of items in the original scale between brackets), example items, scale reliabilities for first and second exposure, and scale sources.

Table 1: Measurement instrument, reliabilities after first round of exposure to Facade (t1) and second round of exposure (t2)

Descriptions and reliabilities of scales employed

Scale	Items (orig.)	Example item	$\alpha\ t_1$	$\alpha\ t_2$	Source
Preconditions (Part A)					
System usability	3(3)	"I thought the system was easy to use"	.86	.83	Adapted from Brooke (1996)
User satisfaction	2(11)	"I expected the experience to be more enjoyable" (N)	.58** (*r*)	.41** (*r*)	Authors
Presence	3(6)	"I felt like I was part of the environment in the presentation"	.76	.82	Wirth et al. (2007)
Character believability	2(2)	"I could feel what the characters in the environment were going through"	.35* (*r*)	.37** (*r*)	Authors
Effectance	2(6)	"My inputs had considerable impact on the events in the story"	.85** (*r*)	.53** (*r*)	Klimmt et al. (2007)
Experiential qualities (Part B)					
Curiosity	3(9)	"During the experience, I felt inquisitive"	.76	.76	Spielberger et al. (1979)
Suspense	2(8)	"At some moments I was anxious to find out what would happen next"	.48** (*r*)	.44** (*r*)	Authors, based on Vorderer et al. (1996)
Flow	5(5)	"During the experience I felt competent enough to meet the demands of the situation"	.71	.66	Jackson et al. (2008)
Aesthetic pleasantness	3(5)	"The experience was inspiring"	.83	.85	Adapted from Rowold (2008) and Cupchik et al. (1994)
Pride	4(4)	"I think I have done a good job in bringing the story forward"	.89	.85	Authors

Enjoyment	2(13)	"The experience was entertaining"	.85** (*r*)	.84** (*r*)	Authors, based on Vorderer et al. (2004)
Specific experience measures (Part C)					
Emotional state: positive	3(10)	"At this moment I feel excited"	.80	.86	Watson et al. (1988)
negative	3(10)	"At this moment I feel sad"	.57	.75	Watson et al. (1988)
Role adoption	2(3)	"During the experience I felt more like the character then like myself"	.41** (*r*)	.62** (*r*)	Authors

Similarly to the previous studies that involved the toolkit scales, most elements of the questionnaire reached satisfying levels of reliability in both rounds of measurement application. Building on this fundamental aspect of scientific quality, the next step of analysis was to determine within-person stability of the measures. This test/retest stability was assessed by inspecting the correlations for each scale value obtained from measurement round 1 and measurement round 2. For example, the ratings on "suspense" that a given participant made after the first exposure to facade was correlated with his/her rating of suspense made after the second round of exposure. Table 2 summarizes the findings from the correlational analysis of test-/retest stability.

Table 2. Correlational analysis of measurement stability across two rounds of interaction with "Facade". "Correlations" mean Pearson' r coefficients of associations between respondents' rating of a user experience dimension after the first round of exposure (T1) and their rating of the same dimension after the second round of exposure (T2).

User experiences	*Correlations T1 / T2*	*p*
A Preconditions		
System usability	.76	<.01
Correspondence /w user expectations	.31	<.05
Presence	.71	<.01
Character believability	.37	<.01
Effectance	.27	=.05
B Experiential qualities		
Curiosity	.39	<.01
Suspense	.24	=.09
Flow	.47	<.01
Aesthetic pleasantness	.59	<.01
Enjoyment	.71	<.01
Pride	.30	<.05
C Specific exp. measures		
Affect: positive	.59	<.01
negative	.57	<.01
Role adoption	.47	<.01

The correlational analysis revealed that some dimensions of the user experience stayed very stable across the two rounds of interaction with "Facade", especially usability, presence, and enjoyment. Most other dimensions displayed considerable stability with Pearson coefficients of .40 and higher. Some dimensions, such as suspense, effectance, and pride, were found to be rather instable across the two measurement applications. All correlations are positive and statistically significant (except for the suspense dimension, which is only close to 5 % significance). The conclusion is that users tried out more different interaction options and story-related inputs during the second round of exposure, as they had become better acquainted to the technical ramifications (see also next paragraph below). Thus, what had been done during the first round of exposure and had triggered some kind of experience then was not necessarily reproduced during the second phase of exposure, so that specific types of experience (such as suspense) tended do shift. The fact that 'holistic', fundamental categories of the experience such as enjoyment and presence displayed especially high stabilities suggests, however, that the actual measures work sufficiently well; we interpret the findings in the way that the instrument is sufficiently stable, but displays factual changes in what users did in round one versus round two and the consequences of these shifted interaction behaviours for the actual experience.

The final step of data analysis was to compare participant responses between the two rounds of exposure to find out whether the repeated interaction with the Facade system shifted the user experience in a systematic, interpretable fashion. Paired t-tests on the average ratings obtained for each of the 13 assessed dimensions of user experience were conducted for this purpose. Table 3 shows the results, including means, standard deviations, and significance of difference between the two rounds of exposure. Findings show that participants experienced significantly higher system usability ($t(49) = 3.57$, $p < .01$), effectance ($t(49) = 2.21$, $p < .05$), and flow ($t(49) = 3.45$, $p < .01$) after the second interaction with Facade than after the first interaction. With marginal significance, the same pattern occurred for Presence ($t(49) = 1.91$, $p < .07$). For the remainder experiential dimensions no differences between first and second exposure were found. A weak tendency emerged, however, in ratings of character believability that went down after the second interaction with the system.

Table 3: Comparison of user ratings between the first and second round of interaction with Facade

	1st exposure		2nd exposure		
User experiences	*M*	*SD*	*M*	*SD*	*P*
Preconditions (Part A)					
System usability	3.66	1.00	3.99	.79	**.001***
Correspondence /w user expectations	2.89	.94	2.99	.80	.46
Presence	3.25	.80	3.41	.75	**.06†**
Character believability	3.68	.61	3.49	.80	.13
Effectance	3.00	1.09	3.41	.97	**.03***
Experiential qualities (Part B)					
Curiosity	3.48	.65	3.46	.78	.87
Suspense	3.53	.91	3.50	.81	.86
Flow	2.86	.69	3.22	.65	**.001***
Aesthetic pleasantness	2.45	.93	2.45	.88	.96
Pride	2.38	.96	2.57	.98	.24
Enjoyment	3.15	1.11	3.20	1.15	.69
Specific experience measures (Part C)					
Affect: positive	2.71	.80	2.76	.91	.64
negative	2.77	.77	2.60	.87	.11
total	2.97	.60	3.08	.70	.19
Role adoption	2.97	.96	2.88	.99	.55

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

Discussion

Findings are interesting both in terms of method development and in terms of studying user reactions to interactive stories from a conceptual point of view. First, results once again demonstrate good statistical performance, as reliabilities of most scales were found to be sufficiently reliable; many subscales achieved even better reliability ratings. Moreover, the results add satisfying test-/retest stability to previous positive judgments of scale quality. Although not all correlations between T1 and T2 measurements were very high, they were all positive and statistically significant. No indication of 'maverick' scales with alarming instability

during repeated applications was found.

From a conceptual point of view, the exploration of how user experiences change with a second exposure to an unfamiliar interactive story returned important insights. Shifts in the user experience were primarily related to issues of interaction and agency: After the second round of interaction, users were more satisfied with the system's usability and experienced higher levels of efficacy (effectance). This implies that respondents found it easier to make the system or story do what they intended to do, and they detected their own impact on the events in the story with greater accurateness and/or immediacy. With smoother interaction that probably resulted from better accommodation to the technical requirements of the "Façade" system, respondents experienced a greater level of flow. This can be interpreted as either a more stable (long-lasting, non-interrupted) flow experience or a greater likelihood of (moments, or phases of) intense flow experiences. Because ratings for Presence were also higher after the second round of interaction, the conclusion is that once users got acquainted to the rules, possibilities, and constraints in interacting with the system, they had a more engaging, immersive experience in terms of active use and action-within-story. However, those experiential dimensions that are closer bound to story development, such as suspense, aesthetic pleasantness, or surprise, did not change between the two rounds of interaction. So the gains in terms of user experience that were detected by the measurement toolkit are focused on the 'interactive' part of interactive storytelling. Acting became easier for participants, whereas the narrative components were not experienced differently.

A secondary finding is the tendency towards lower levels of character believability after the second round of exposure, which may be read as an indication of users trying out more different ideas during the second round. Such more exotic or, from a system design perspective, unconventional user actions may have made the limits in the system's artificial intelligence and characters' responsiveness more salient to users than during the first round of interaction. Consequently, character believability went down. An analysis of the user input during round one and round two may illuminate this issue in future work.

Overall, the second experiment that involved "Façade" once again brought out positive outcomes in terms of measurement development. The scales have now been demonstrated to be reliable several times, including a test/retest approach in the current study. Moreover, some interesting findings have been obtained from comparing user experiences after the first and the second round of interaction, which suggest a rather stable perception of the overarching narrative, but shifts in users' experience of their own agency and control of the system. The fact that both the stability analysis and the inspection of mean differences revealed changes in user behavior and experience from one round of system use to the next indicates, however, that researchers who apply toolkit measures (and any other measures of user experience as well) need to consider the impact of exposure time, user experience with the system, and training phases when planning user-centered evaluation studies.

2.2 Application of the Instrument to a Contemporary Prototype: "EmoEmma"

Background and Scope

With increasing evidence for the statistical quality, practicability, and robustness of the toolkit scales on user experiences, an important addition to the collection of empirical studies within WP7 is to expand the list of IS systems and prototypes from which reference data are obtained. Previous studies had confronted end-users with the video game "Fahrenheit" (D7.2)

and the widely known foundational prototype “Facade” (D7.2, and section 2.1. of the current report). The present study was intended to add a state-of-the-art piece of IS technology to the portfolio of WP7 research, namely the “EmoEmma” system that IRIS partner TEES has been working on (figure 1). EmoEmma is a system designed for immersive, VR-based interactive storytelling that employs elements from the Madam Bovary novel as narrative content (Cavazza, Pizzi, Lugrin & Charles, 2007). It thus comes with a profoundly different look and feel from both “Fahrenheit”, “Facade”, and other systems of interest. With two different modes of interaction, the “ghost” mode versus the “actor” mode, the system also allows to conduct an in-depth analysis of user responses to technologically important variations in the design of interactive storytelling systems (see method section for details). Moreover, the application of the toolkit scales to this system also allows comparative conclusions, as some user studies have already been conducted with “EmoEmma” (Lugrin, Cavazza, Pizzi, Voigt & Andre, 2010). So expanding the gallery of reference data on user experiences (measured by the toolkit scales) with “EmoEmma” also enabled inspections of convergent validity with previous user research and, by comparing user reactions to “ghost versus actor” mode, also conceptually promising experimental insights as well.

Figure 1. Screenshot from the EmoEmma system by IRIS partner TEES.

Method

An experiment was conducted to compare (1) participants’ responses to actor vs. ghost mode, and (2) (similar to the prior study on Façade) participants’ experiences following first and second exposure to EmoEmma. Both comparisons were made within-subjects. The order in which participants interacted in actor of ghost mode was balanced, which enables us to test for possible order effects. A total of 34 university students (11 males, 23 females; average age M = 22.0 years, SD=1.92 years) with a low to moderate degree of computer game literacy (M=1.71, SD=.84 on a scale from 1 to 3) participated in the study.

Upon arrival in the laboratory, participants received a short training in interacting with EmoEmma for about 5 minutes. Next, half of the participants were first exposed to an IS

sequence in actor mode, whereas the other half was first exposed to ghost mode. Sequence duration was between 4 and 10 minutes, depending on participants' ability to maintain a pleasant conversation with the character of Madame Bovary. Next, participants completed the IRIS toolkit questionnaire including demographical questions, and questions relating to the 14 user experience dimensions in the following order: Curiosity, Suspense, Flow, Aesthetic pleasantness, Enjoyment, Affect, Role adoption, System usability, User satisfaction, Character believability, Effectance, Presence, Autonomy, Pride. User experience dimensions were measured using between 2 and 5 items each. Subsequently, participants proceeded to interact with EmoEmma, but now in the opposite (ghost or actor) mode. Then they completed the IRIS toolkit questionnaire for the second time. Logs of both interactions with EmoEmma were recorded and saved. Upon completion of the second questionnaire, participants received 20 EUR as compensation, and were debriefed and dismissed.

Results

First, we tested the reliability of all 14 scales measuring user experience dimensions. We used α = .70 as the default benchmark of sufficient reliability, although α = .60 were determined as minimum requirement.

Reliability analyses show almost all scales of the standardized assessment tool met the minimal requirement following both rounds of exposure, with α values ranging between .61 and .88 (N = 34). The only exception is the reliability for suspense at both first (α = .56) and second (α = .56) exposure. Also, we had to remove one item from the negative affect subscale scale to optimize reliability. Reliabilities of the two-item scales of effectance, suspense, enjoyment, and role adoption were assessed using a Pearson's correlation, and were evaluated as sufficient. For the 2-item scales of user satisfaction and character believability, reliabilities at second exposure were insufficient – for both constructs, we used only 1 item. Table 4 summarizes the measurement instruments used, including numbers of items, example items, scale reliabilities for first and second exposure, and scale sources.

Table 4: Measurement instrument, reliabilities after first round of exposure to EmoEmma (t1) and second round of exposure (t2)

Descriptions and reliabilities of scales employed

Scale	Items	Example item	*Reliability Alpha t_1*	*Reliability Alpha t_2*
Preconditions (Part A)				
System usability	3	"I thought the system was easy to use"	.61	.70
User satisfaction	1	"The experience was better than I expected"	n/a	n/a
Presence	3	"I felt like I was part of the environment in the presentation"	.79	.76
Character believability	1	"I could feel what the characters in the environment were going through"	n/a	n/a

Effectance	2	"My inputs had considerable impact on the events in the story"	.82** (*r*)	.88** (*r*)
Autonomy	4	"I noticed many opportunities to influence the story"	.84	.87
Experiential qualities (Part B)				
Curiosity	3	"During the experience, I felt inquisitive"	.72	.82
Suspense	4	"At some moments I was anxious to find out what would happen next"	.56	.56
Flow	5	"During the experience I felt competent enough to meet the demands of the situation"	.62	.71
Aesthetic pleasantness	3	"The experience was inspiring"	.81	.87
Pride	4	"I think I have done a good job in bringing the story forward"	.87	.88
Enjoyment	2(13)	"The experience was entertaining"	.86** (*r*)	.94** (*r*)
Specific experience measures (Part C)				
Emotional state: positive	3	"At this moment I feel excited"	.86	.84
negative	2	"At this moment I feel sad"	.36* (*r*)	.37* (*r*)
Role adoption	1	"I felt like I was in the main characters skin"	n/a	n/a

In sum, most elements of the questionnaire reached satisfying levels of reliability in both rounds of measurement application. As a next step we inspected the correlations for each scale value obtained from first and second exposure. Table 5 summarizes the findings from the correlational analysis of test-/retest stability.

Table 5. Correlational analysis of measurement stability across two rounds of interaction with "EmoEmma". "Correlations" mean Pearson' r coefficients of associations between respondents' rating of a user experience dimension after the first round of exposure (T1) and their rating of the same dimension after the second round of exposure (T2).

User experiences	*Correlation T1 / T2 (Pearson's r)*	*p*
A Preconditions		
System usability	.67	<.01
Correspondence /w user expectations	.62	<.01
Presence	.65	<.01
Character believability	.08	n.s.
Effectance	.26	n.s.
Autonomy	.57	<.01
B Experiential qualities		
Curiosity	.31	=.08
Suspense	.53	<.01
Flow	.53	<.01
Aesthetic pleasantness	.77	<.01
Enjoyment	.73	<.01
Pride	.05	n.s.
C Specific exp. measures		
Affect: positive	.72	<.01
negative	.76	<.01
Role adoption	.73	<.01

Like in the Facade study (see 2.1), correlational analyses revealed that most dimensions of user experience stayed quite stable or even very stable across the two rounds of interaction with EmoEmma (coefficients of .40 and higher). Also like in the prebious Facade study, effectance and pride were rather instable across the two measurement applications – for EmoEmma this instability also holds for character believability and, to a lesser extent, curiosity. This means that on these measures, the quality of the first experience does not predict the quality of the second experience very well: Participants who had a positive first experience may have had a more negative experience on second exposure, and vice versa. Possibly, this finding could be related to the order in which participants were exposed to actor and ghost mode. We will address this issue later. First, however, we will analyze main effects of (1) exposure to actor vs. ghost mode, and (2) first vs. second exposure to EmoEmma.

To analyze whether user experiences differed for exposure to actor vs ghost mode, we conducted a series of paired t-tests. Results, including means, standard deviations, and significance of difference between both modes are shown in Table 6.

Findings show a fairly consistent preference for the ghost mode in participants. Participants experienced significantly higher effectance ($t(33) = 2.88$, $p < .01$), more pride ($t(33) = 3.73$, $p < .005$), and more positive total affect ($t(33) = 4.72$, $p < .001$) in ghost mode than in actor mode. In addition, they experienced marginally higher satisfaction ($t(33) = 1.79$, $p < .09$), autonomy ($t(33) = 2.03$, $p < .06$), curiosity ($t(33) = 2.02$, $p < .06$), and flow ($t(33) = 1.70$, $p < .1$) in ghost mode. For the remainder experiential dimensions no significant differences between actor and ghost mode were found.

Table 6: Comparison of user ratings between EmoEmma's Actor and Ghost mode

	Actor mode		Ghost mode		
User experiences	*M*	*SD*	*M*	*SD*	*P*
Preconditions (Part A)					
System usability	4.11	.80	4.22	.65	.28
Correspondence /w user expectations	3.09	1.06	3.35	.98	**.08†**
Presence	3.26	.85	3.10	1.01	.22
Character believability	3.12	.98	3.06	1.04	.80
Effectance	2.24	.95	2.88	1.27	**.007***
Autonomy	2.17	.85	2.47	1.08	**.05†**
Experiential qualities (Part B)					
Curiosity	3.59	.80	3.86	.56	**.05†**
Suspense	3.61	.61	3.49	.80	.37
Flow	3.09	.71	3.31	.79	**.09†**
Aesthetic pleasantness	2.33	.91	2.44	.94	.34
Pride	2.21	.84	3.01	1.10	**.001***
Enjoyment	3.68	1.10	3.69	.95	.91
Specific experience measures (Part C)					
Affect: positive	3.00	.95	3.17	.91	.14
negative	1.97	.82	1.88	.65	.33
total	3.01	1.01	3.65	.55	**.000***
Role adoption	2.76	1.14	2.63	1.02	.34

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

In sum, EmoEmma's ghost mode appears to perform better in terms of preconditions (satisfaction, effectance, autonomy), experiential qualities (curiosity and flow), and specific experience measures (affect).

To analyze whether user experiences changed from first to second exposure to EmoEmma, we again conducted a series of paired t-tests. Table 7 shows the results, including means, standard deviations, and significance of difference between the two rounds of exposure.

Findings show that participants experienced significantly higher suspense ($t(33) = 3.02$, $p < .005$) after the second interaction with EmoEmma than after the first interaction. In contrast, participants experienced marginally lower effectance ($t(33) = 1.84$, $p = .07$) and enjoyment ($t(33) = 1.71$, $p = .09$) after second exposure. For the remainder experiential dimensions no significant differences between first and second exposure were found. Surprisingly, this was also the case for system usability. Possibly, the training session preceding the actual experiment improved system usability for the first exposure, and thus dampened a possible learning effect. Moreover, in contrast to the Facade study (section 2.1.), second exposure to EmoEmma implied some changes in the interaction (either a shift from ghost mode to actor mode or vice versa), which has certainly affected user ratings (see table 6 again) and may thus overshadow potential accommodation effects that occurred with interactivity-related variables in the Facade study.

Table 7: Comparison of user ratings between the first and second round of interaction with EmoEmma

	1st exposure		2nd exposure		
User experiences	*M*	*SD*	*M*	*SD*	*p*
Preconditions (Part A)					
System usability	4.24	.61	4.09	.83	.15
Correspondence /w user expectations	3.21	.95	3.24	1.10	.85
Presence	3.27	.92	3.09	.95	.17
Character believability	3.09	1.00	3.09	1.03	1.00
Effectance	2.78	1.14	2.34	1.15	**.07†**
Autonomy	2.36	.90	2.27	1.06	.58
Experiential qualities (Part B)					
Curiosity	3.77	.61	3.68	.79	.50
Suspense	3.37	.71	3.73	.67	**.003***
Flow	3.11	.71	3.29	.80	.17
Aesthetic pleasantness	2.32	.87	2.45	.98	.26
Pride	2.79	.97	2.43	1.12	.16
Enjoyment	3.79	.98	3.57	1.06	**.09†**

Specific experience measures (Part C)

Affect:	positive	3.06	.93	3.12	.94	.63
	negative	1.90	.70	1.96	.77	.51
	total	3.58	.63	3.58	.61	1.00
	Role adoption	2.65	1.10	2.75	1.06	.46

So user experiences may not only be influenced by repeated exposure, or by story mode, but also by the order in which interaction with the two different story modes took place. It could be that for first time users, ghost mode is a somewhat complex experience. Starting in actor mode would facilitate a more natural “build up” in experiences. In addition, from the analysis on actor vs ghost mode we learned that experiences with ghost mode were on average better. Therefore, it could be expected that exposure to actor mode after ghost mode is a somewhat disappointing experience. Therefore, participants who atarted out with actor mode may have had have overall better experiences that those who started out with actor mode.

To test for order effects, we conducted a series of GLM repeated measures analyses. These analyses indeed showed an effect of the order of exposure for curiosity ($F1, 32) = 7.63$, $p < .01$), role adoption ($F1, 32) = 3.62$, $p < .07$ - marginal), user satisfaction ($F1, 32) = 7.54$, $p < .05$), and system usability ($F1, 32) = 7.54$, $p < .05$). In all cases, user experiences were more positive when participants started out interacting in the actor mode.

Discussion

The current study attests to the usability and meaningfulness of the IS questionnaire toolkit to examine responses to IS prototypes under development, such as EmoEmma. Although not all subscales of the toolkit stood up to the .70 benchmark, most turned again out to perform in a satisfying way. For a set extremely brief scales (2 to 5 items each), this is a good result, which shows that it is feasible to measure complex, multi-dimensional user experiences to IS environments in a very concise way. The suspense subscale, as well as two 2-item subscales, did not reach the minimal reliability required, which calls for a reconsideration of the items used.

Surprisingly, the test re-test reliabilities of some scales (effectance, pride, character believability, and curiosity) were rather low. This may be due to the finding that participants who used actor mode first, later had a more positive experience in ghost mode, whereas participants who used ghost mode first, later had a more negative experience in actor mode. For character believability, we conclude that, given its poor reliability scores over several studies, there is still a need for improvement with this particular element of the toolkit.

From our comparison of actor versus ghost mode, we can conclude that, overall, participants had more positive experiences using *ghost* mode. Perhaps participants were less concerned with potential negative responses of Emma to Rodolphe in ghost mode. As one participant put it: ”It was easier to play the ghost, because giving Rodolphe tips about what to say so her, was easier for me than actually say these things to Emma in a convincing way.” Though standing at the sideline may be comfortable, it may also have induced a feeling of lack of control over the situation. As the same participant put it: “However, when being Rodolphe, Emma responded more the way I wanted and when I was playing the ghost, she did not at all what I told her to do. That didn't give me the feeling that I really had a role in the game and

controlled the situation." This latter experience may explain why, in general, participants liked the build up of interacting in actor mode first and in ghost mode later better: Interacting in ghost mode the first time around may have been too much of a challenge.

The results on comparing the actor and ghost mode experience converge with observations reported by Lugrin et al. (2010) on user responses to a highly immersive, 3D version of EmoEmma. In this usability study, participants perceived and exploited the greater degrees of freedom that the ghost mode provides (i.e., a lower need to stay close to the characters, and the chance to move around freely to explore the environment or search for relevant objects). The present results on higher autonomy and curiosity in ghost mode (table 6) correspond to these findings and suggest particular experiential profiles are bound to different styles of interactive storytelling. Maybe due to the (technical) limitations that still come with the actor mode, the ghost mode that is less susceptible to severe interruptions of dialogue and story progress is experienced by users as more appealing.

Finally, we replicated the design of the abovementioned Facade study by comparing first and second exposure to EmoEmma. Here, a surprising result was that suspense was higher during the second exposure, while at the same time participants became a little frustrated with their lack of effectance (marginally lower), which led to marginally lower enjoyment. On the positive side, system usability did not change from first to second exposure, indicating that the system – after some initial training - was rather user friendly. Still, this comparison of two rounds of exposure needs to be interpreted against the fact that in contrast to the Facade study, this time first and second round of exposure did not feature exactly the same interactive storytelling environment, but differed for each participant to a notable degree.

2.3 User Responses to Dialogue-Based Interactive Stories: Application of the Toolkit Instrument to the "Virtual Beergarden"

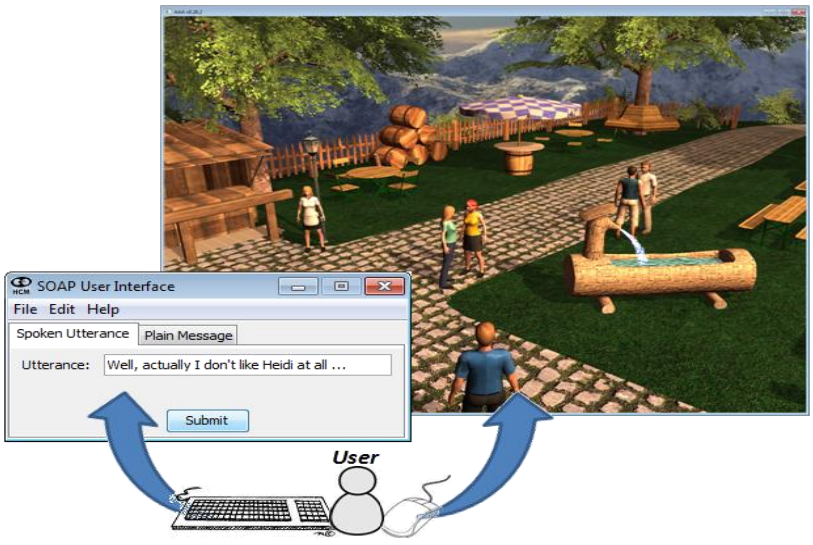
Background and Scope

The research interested pursued by IRIS WP7 closely connects to issues of how to design interfaces and interaction processes between users and Interactive Storytelling systems, which are core topics in IRIS WP6 headed by partner UOA. Following reviewer recommendations and recent developments in WP6, user reactions to specific design decisions concerning the mode of dialogic interaction with the "Virtual Beergarden" prototype developed by UOA were addressed as a case for joint WP6/WP7 research based on cooperation between UOA and VUA. The "Virtual Beergarden" (VB, Figure 2) serves as demonstrator for how conversation-based, dialogic interactions of end-users with an interactive story can be structured and managed from a technological point of view. The setting of this prototype moves away from previous approaches with high-art literature (e.g., the Merchant of Venice) towards mainstream entertainment contents such as soaps and television series. Virtual agents with anthropomorphic appearance (i.e., a waitress and various guests in a beergarden) are displayed in group-wise conversations, and the user can direct his/her character to join these conversation groups and enter statement contributions through the keyboard. User statements are just like agent statements and responses generated with a state-of-the-art text-to-speech processing system so that interactive audio conversations emerge. As a comparatively 'early' prototype, the VB is not a full-scale, narratively complete entertainment medium, but rather a short, exploratory confrontation with a language-based interactive environment. In this sense, the VB is somewhat comparable with "Facade" (from a user interaction perspective), yet very different in terms of content and audiovisual representation. So the VB adds to the portfolio of systems to which the toolkit

measures are applied due to some design features that the previous systems did not have. At the same time, experimental variations in system design once again may allow drawing conceptual conclusions on user experiences with (in this case: dialogue-based) interactive 'soap' stories.

One relevant difference in designing dialogue-based interaction with a digital story refers to round-based versus continuous organization of conversation flows between users and agents. In a round-based dialogue organization, users can only participate in conversations and thus affect story progress within certain time windows that are opened and closed by the system. Users have to 'wait' until the virtual agents have finished their statements; only afterwards users can type in their utterance and make agents react to it. As a result, the flow of conversation is 'clean', because no interruptions can occur, and agents only need to respond to single, clearly defined user inputs. The step-by-step organization of dialogue thus enables the system to produce optimal reactions to user inputs. But the price for these advantages is that for users, the process of conversation may appear to be not very natural, authentic, and intuitive. Having to wait until one's time for a statement has come may – depending on the narrative context and the current conversation topic – undermine users' sense of immersion and presence, and may also affect perceived character believability. On the other hand, the alternative design strategy of continuous dialogue options, which enables users to make statements any time and also during utterances made by virtual agents, may cause serious confusion for the language processor and the user alike, as ill-timed contributions to the conversation may lead to less-than-optimal agent responses and incidents of communication chaos among virtual and human participants of the dialogue. Users may also misuse the option to affect conversation any time to empower themselves over the extent that their role (in the present case: their role as another guest of the beergarden) would suggest; by overriding agent contributions to the ongoing conversations, the whole virtual-social experience may be damaged severely. In this sense, a round-based dialogue would force users to 'take the agents serious' and listen to their statements. In sum, there are conceptual arguments why either round-based or continuous modes of dialogic interaction between the user and the virtual agents may result in more or less favourable user experiences and overall enjoyment. Thus, the case of designing for continuous versus round-based interaction was chosen for an experimental test of the VB involving end-users and the measurement toolkit scales.

Figure 2. Screenshot from the "Virtual Beergarden" (by IRIS partner UOA)



Method

A total of 42 university students (mean age: 22 years, 30 females) participated in the study. They were invited to express their opinions about a new type of entertainment software, for which they would be shown a short demo version. After they received a brief introduction to the VB, especially concerning the dialogue interface, they interacted with the VB for five minutes; when the time was over, the experimenter kindly interrupted the ongoing interaction and presented a questionnaire that included the self-report scales (short version as in the second "Facade" study, see section 2.1). Participants were randomly assigned to either use a version of the VB with continuous dialogue organization or a version with round-based conversation. Except for this variation of dialogue progress, both versions of the VB were identical – the agents were the same, and narrative content was held constant across conditions. Users either controlled a female or a male avatar, according to their own sex. The technical specifications of the VB system have been reported by IRIS partner UOA in deliverable D6.4. Upon completion of the questionnaire, participants were confronted with a different system, which served as data collection for the final IRIS WP7 validation and benchmarking experiment (see section 2.4.).

Results

As with previous studies, the priority of data analysis was assigned to scales' reliability. Findings show once again mostly very good or at least satisfying scale homogeneity, which is of notable value because the scales were very short (table 8). Again, the character believability scale turned out to perform suboptimally, and also the 'very short flow scale' with only three items failed to reach acceptable Alpha values. All other parts of the toolkit performed well.

Table 8. Reliabilities of Toolkit scales in the "Virtual Beergarden" study

Scale	Items (orig.)	Example item	*Cronbach's α*
Preconditions (Part A)			
System usability	3(3)	"I thought the system was easy to use"	.85
User satisfaction	3(11)	"I was satisfied with my use of the system"	.65
Presence		Not assessed in Beergarden Study	
Character believability	2(2)	"I could feel what the characters in the environment were going through"	.39* (*r*)
Effectance	2(6)	"My inputs had considerable impact on the events in the story"	.89 (*r=.82***)
Autonomy (new scale)	4 new items	"I had the impression that I could make many different events happen in the story"	.85
Experiential qualities (Part B)			

Curiosity	3(9)	"During the experience, I felt inquisitive"	.79
Suspense	4(8)	"At some moments I was anxious to find out what would happen next"	.71
Flow	3(5)	"During the experience I fully concentrated on my task"	.46
Aesthetic pleasantness	3(5)	"The experience was inspiring"	.82
Pride		Not assessed in Beergarden Study	
Enjoyment	2(13)	"The experience was entertaining"	.90 (*r* =.82**)
Specific experience measures (Part C)			
Emotional state: positive	9(10)	"Now, after the experience, I feel enthusiastic"	.87
negative	10(10)	"Now, after the experience, I feel nervous"	.87
Role adoption	3(3)	"During the experience I felt more like the character then like myself"	.71

In addition to the reliability analysis, group differences between participants who had used the continuous dialogue version of the virtual beergarden and respondents who had been confronted with the round-based dialogue version were inspected. Given that the technical manipulation was relatively 'small' (i.e., no fundamental changes to the overall system had been made), it was not expected to obtain massive group differences across many of the examined dimensions of the user experience. In line with this modest expectation, most toolkit measures did not return a significant mean difference between continuous and round-based dialogue conditions (table 9). Yet meaningful differences emerged at the new dimension "autonomy" (see section 2.4 for conceptual details) and curiosity; other fun-related aspects, such as suspense and enjoyment displayed similar yet non-significant tendencies.

Table 9. Experimental group comparison of user experiences in the "Virtual Beergarden" study

	Round-based dialogue		Continuous dialogue		
User experiences	*M*	*SD*	*M*	*SD*	*p*
Preconditions (Part A)					
System usability	3.97	1.01	3.82	1.11	.64
Correspondence /w user expectations	3.50	1.44	3.40	1.19	.81
Presence	not assessed in Beergarden study				
Character believability	2.91	0.83	2.58	0.96	.23

Effectance	2.86	0.85	3.08	1.20	.51
Autonomy	2.17	0.79	2.68	0.93	**.055**

Experiential qualities (Part B)

Curiosity	3.33	0.92	3.78	0.77	**.095**
Suspense	2.82	0.79	3.16	0.74	.15
Flow: Single Item*	4.09	0.68	4.30	0.57	.29
Aesthetic pleasantness	1.62	0.58	1.60	0.92	.92
Pride	Not assessed in Beergaden Study				
Enjoyment	3.07	0.94	3.45	0.95	.20

Specific experience measures (Part C)

Affect: positive	2.38	0.75	2.20	0.83	.53
negative	1.32	0.29	1.46	0.75	.43
Role adoption	1.83	0.65	2.07	0.95	.36

*Due to low scale reliability, the single item "I was completely focused on the task at hand" was used as proxy for flow experience in this analysis.

Discussion

The application of the toolkit measures to partner UOA's "virtual beergarden" has once more stabilized the performance metrics in terms of reliability. The character believability scale emerges as key access point for further method improvement; this task remains for future studies, also beyond the IRIS timeframe and work plan. With only three items, the "flow" scale has obviously been shortened to intensively; a return to longer versions is indicated for future studies, as a five-item scale had been applied successfully in previous studies (section 2.2.).

The experimental design of the "Beergarden" study revealed that the toolkit's sensitivity for the impact of relatively small technical system changes on user responses is satisfying. While the great majority of system features was held constant across the two experimental conditions, the dialogue management – as one 'minor' component of the overall system – differed systematically, yet not in a profound way: Either users could type in their dialogue contributions at any time they wanted (continuous dialogue), or they had to wait each time until they were enable to 'speak' / type. The fact that the scales returned notable group differences on the dimension of autonomy (see next section: 2.4) and curiosity is both interesting in terms of conceptual and technology design perspectives (see D6.4 as well as a related submission to the ICIDS 2011 conference: Endrass et al., 2011) and in terms of scale performance. Making the user wait versus allowing her/him to decide about the timing of her/his contribution autonomously is conceptually linked to perceived autonomy. This is nicely

reflected by the results. Similarly, the impact of dialogue style on curiosity is interpretable in the sense that a continuous dialogue system is more likely than a round-based dialogue system to generate the feeling of 'anything can happen' and of greater system responsivity to spontaneous dialogue contributions. So overall, the experimental findings make an important case for the toolkit's measurement qualities, because the instrument turns out to be useful in the user-centered evaluation of 'early prototype' IS systems or demonstrators for particular elements of complete IS systems such as the "Beergarden".

2.4 Autonomy as an Experience of Interactive Stories: Application of the Instrument to "Fable: The Lost Chapters"

The final element of IRIS WP7 studies on scale development and validation again served two purposes: One was the expansion of the list of investigated IS systems by another, partially unique approach; the other was to address an issue that was found to be conceptually highly relevant, yet underrepresented in the previous theoretical frameworks and in the composition of the toolkit. This theoretical addition refers to *autonomy* as a dimension of user experience. So far, the conceptual framework underlying the toolkit measures includes several reflections of the assumption that interactive use will shape the user experience. The concept of effectance is bound to users perceiving their impact onto the system and the story; the concept of flow (Csikszentmihalyi, 1990) refers to the immersive quality of challenging-yet-manageable interaction (Klimmt, Roth, Vermeulen, Vorderer & Roth, in press). The dimension of pride was added as conceptual reaction to users understanding of narrative progress as personal success achieved through effective, intelligent input (Klimmt, Vorderer, & Nuss, 2010). Following recent literature in video games research, however, one conceptually distinct facet in users' perception and experience of interactivity in entertainment contexts was found to be equally relevant: autonomy. Regardless of whether users experience themselves as effective or victorious, or of whether they enter a state of flow during interaction with a story or not, users may value the broadness of options they are allowed to decide on. A well-established theoretical framework of human motivation, self-determination theory (Deci & Ryan, 2000), emphasizes people's need and desire for experiencing autonomy. Several studies on video games that build on this framework found that playing digital games fulfils the need for autonomy quite effectively (Ryan, Rigby, & Przybylski, 2006; Tamborini, Bowman, Eden, Grizzard & Organ, 2010).

In the context of Interactive Storytelling, the experience of autonomy relates to various aspects of user interaction with the technical system and with the narrative content. Users should perceive relatively high (low) degrees of autonomy if they are enabled to give input to the technical systems through several (few) channels, if the story progress can be manipulated at many (few) occasions such as branching points, and if users have the impression that they can shape the ongoing story towards various diverse (only a few and similar) events and outcomes. Thus, the user experience of autonomy not only closely links to overall media enjoyment (Tamborini et al., 2010), but also mirrors important trends in IS technology development – many innovations, for instance, in dynamic story generation and planning – are intended to enlarge users' objective autonomy. To the extent that such innovations in objective autonomy also translate into subjective experiences of autonomy, the measurement toolkit should be able to monitor this pathway to meaningful and enjoyable experiences. Therefore, perceived autonomy was added to the list of concepts included in the toolkit instrument.

The present study served to generate empirical and conceptually relevant insight with particular focus on the new dimension of autonomy. In addition to application of the actual measure of autonomy (see next section for details), the research design also manipulated the

users autonomy during exposure to the role playing game “Fable: Lost Chapters” (Lionhead Studios / Microsoft, 2005, see figure 3). This game pursued a philosophy of strong player autonomy and highly consequential player decisions; it thus immediately addressed the autonomy dimension of user experiences. Moreover, as a narratively dense video game, it represents another mode of interactive storytelling. As an off-the-shelf product with completed story and cinematic graphics, it mirrors the product characteristics of “Fahrenheit”, which had been examined in one of the first WP7 studies (D7.2). An innovative experimental setup was implemented to study user reactions to high levels versus low levels of story-related autonomy. This way, a conceptual aspect of interactive storytelling design was addressed at the same time as the IRIS WP7 portfolio of examined IS prototypes and systems was expanded.

Figure 3. Screenshot from the introduction sequence of “Fable: Lost Chapters”

Method

The study was implemented as second part of the laboratory session that the participants of the “Virtual Beergarden” study (see section 2.3.) went through (N = 42, average age = 22 years, 30 females). After the exposure to the “Beergarden” and completing the according questionnaire, participants were invited to play the introductory sequence of “Fable: Lost Chapters”. In this sequence, which takes on average about 20 to 30 minutes to complete, players are made familiar with the overarching narrative and are confronted to resolve four tasks that involve moral decisions. These decisions – for example, (1) accepting money from a man who cheated his wife for not telling the wife versus (2) rejecting the money and telling the wife versus (3) accepting the money but still telling the wife, serve to define the player character’s moral personality, which has long-term impact for story development. In the present study, the moral decisions required from players were used to manipulate players’ autonomy.

For this purpose, one half of the participants was instructed to play the introduction sequence in the normal way; that is, they were free to make all requested moral decisions according to their personal will (condition: autonomy = high). The opposite half of the participants,

however, was not allowed to make such autonomous decisions, but was given a list of instructions on how to decide in each of the moral tasks in the game. This way, players had to click and implement decisions, however, the content of the decisions had been given from an external authority (i.e., the experimenter) and was thus not the product of player's free, autonomous choice. So even if the instructed decision was morally agreeable for the participant, s/he could not perceive herself or himself as the originator of the decision (condition = low).

This procedure threatened experimental internal validity, because players with high autonomy might have generated a broader diversity in story progress than the group with low autonomy. To rule out such so-called confoundings, an innovative experimental procedure was implemented that follows recommendations by Klimmt, Vorderer and Ritterfeld (2007). Participants were randomly assigned to the experimental conditions by sorting the first participant on the schedule list to the high autonomy condition. His/her decisions on the five moral tasks in the game were recorded. The second participant on the list was assigned to the low autonomy condition and received instructions for the moral tasks that exactly replicated those decisions that the first participant had made autonomously. Participant number three was again assigned to the high autonomy condition and was thus allowed to make free-will decisions in the game; his/her decisions then served as protocol for what the fourth participant (again in the low autonomy condition) was instructed to do. Overall, this procedure ensured that the content of decisions and the game was held constant across experimental conditions. The only difference between the two experimental groups was that one group of players had made all the decisions by their free will, whereas the other groups typed in the very same decisions based on external instruction. Thus, a very 'clean' experimental variation of user autonomy was implemented.

After completing the introduction sequence of "Fable: Lost Chapters", participants filled in the measurement toolkit questionnaire. It included a new short scale for perceived autonomy, with the following items (English translations; original wording was in German):

- I had the impression that I could make many different events happen in the story.
- I perceived many possibilities to influence the progress of the story.
- I experienced strong limitations to my decisions about the progress or the story (reverse-coded).
- The system allowed me to affect the story exactly in the way I had in mind.

It was assumed that this autonomy scale – if reliable and valid – should respond directly to the experimental manipulation of user autonomy. Moreover, the design and the application of the autonomy measure within the toolkit instrument allowed to examine whether and how autonomy as a user experience relates to overall enjoyment and to other facets of interactivity-based experience such as effectance.

Results

The reliability analysis of the "Fable" study completes the overall positive picture of the instruments' homogeneity that had been concluded from the previous experiments. With the exception of character believability, all scales met the minimum requirements for Cronbach's Alpha, and even the character believability items performed better than in the preceding studies (2.2. and 2.3.). Table 10 summarizes the reliability findings.

Table 10. Reliabilities of Toolkit scales in the "Fable: Lost Chapters" study

Scale	Items (orig.)	Example item	*Cronbach's α*
Preconditions (Part A)			
System usability	3(3)	"I thought the system was easy to use"	.89
User satisfaction	3(11)	"I was satisfied with my use of the system"	.69
Presence		Not assessed in Fable Study	
Character believability	2(2)	"I could feel what the characters in the environment were going through"	.53 (r=.37)
Effectance	2(6)	"My inputs had considerable impact on the events in the story"	.90 (r = .82)
Autonomy (new scale)	4	"I had the impression that I could make many different events happen in the story"	.86
Experiential qualities (Part B)			
Curiosity	3(9)	"During the experience, I felt inquisitive"	.84
Suspense	4(8)	"At some moments I was anxious to find out what would happen next"	.72
Flow	4(5)	"During the experience I fully concentrated on my task"	.71
Aesthetic pleasantness	3(5)	"The experience was inspiring"	.83
Pride		Not assessed in Fable Study	
Enjoyment	2(13)	"The experience was entertaining"	.94 (r =.89)
Specific experience measures (Part C)			
Emotional state: positive	9(10)	"Now, after the experience, I feel enthusiastic"	.87
negative	10(10)	"Now, after the experience, I feel nervous"	.96
Role adoption	3(3)	"During the experience I felt more like the character then like myself"	.80

The second interest in the data was again the experimental approach, that is, the comparison of user experiences in players who had made free decisions during game play (autonomy = high) and this participants who had been instructed to make particular, pre-defined choices. An analysis of the decisions made by participants in the high autonomy group revealed that the vast majority of choices was morally positive. For two of the four decisions, all participants who were allowed to make a free choice selected the morally favourable option. For the other two decisions, only very few players decided to depart from the morally most positive options. Therefore, the moral implications of those decisions that 'low autonomy' participants were instructed to make according to a task list were virtually always positive. This has probably reduced the perception of reduced autonomy substantially, because most 'forced' decisions did not come with a feeling of moral wrongness. If instructions to low-autonomy players had involved a higher frequency of 'having to do what is morally inappropriate', the perceived minimization of autonomy would probably have turned out stronger. So the analysis of manipulation strength of autonomy that was achieved with the present design reveals a rather small discrepancy in the conditions set up for high autonomy versus low autonomy participants. Consequently, no fundamental discrepancies in user experiences were expected.

The findings confirm the assumption of low experiential differences between more and less autonomous participants (table 11). None of the 13 experience measures applied in the study returned significant or near-significant group differences. Specifically, the autonomy scale that was expected to react most directly to the experimental manipulation of players' freedom to decide did not reveal any substantial group difference. The only dimension where a systematic tendency 'began to emerge' was role adoption or identification, where the greater freedom to decide about one's character's personality led to modestly increased average values. So while the scales' reliabilities were again satisfying, the present experimental study element did not produce notable effects in the sense of systematic group differences..

Table 11. Experimental group comparison of user experiences in the "Fable: Lost Chapters" study

	Forced user decisions (autonomy = low)		Free user decisions (autonomy = high)		
User experiences	*M*	*SD*	*M*	*SD*	*p*
Preconditions (Part A)					
System usability	3.76	1.02	3.78	1.06	.60
Correspondence /w user expectations	3.86	0.65	3.60	0.93	.31
Presence	not assessed in Fable study				
Character believability	2.95	0.69	3.00	0.97	.86
Effectance	3.40	0.83	3.10	1.25	.35
Autonomy	3.35	0.79	3.02	1.02	.26
Experiential qualities (Part B)					

Curiosity	3.90	0.64	3.86	0.91	.85
Suspense	3.80	0.72	3.57	0.75	.32
Flow	3.54	0.74	3.65	0.74	.60
Aesthetic pleasantness	1.48	0.84	1.75	0.65	.25
Pride	Not assessed in Fable Study				
Enjoyment	3.90	0.90	3.79	1.09	.70

Specific experience measures (Part C)

Affect: positive	2.40	0.91	2.62	0.79	.41
negative	1.38	0.93	1.45	0.61	.76
Role adoption	2.35	1.05	2.83	1.01	.14

Discussion

The final study of VUA's experimental research programme within WP7 contributed to the positive picture to the toolkit's scale performance. Reliability was satisfying to very good. The fact that no experimental group differences emerged needs to be discussed with regard to whether this is an indication of dissatisfying scale sensitivity or a consequence of insufficiently strong experimental manipulation of the game situation and the interactive narrative.

Compared to the "Beergarden" experiment (section 2.3), the present manipulation of the interactive story was not as present for users throughout the experience. Instead, the manipulation was only related to four important action points in the game environment. How participants moved through the story world and explored its geography, as well as the sequence in which they went through the moral decision tasks was not affected by the experimental variation of autonomy. So in the 'low autonomy' group, there was still a lot of free choice left in the sense that players could wander around freely and decide on when they would engage in one of the tasks for which forced-choice instructions had been given. Moreover, virtually all decisions that the high-autonomy participants had made and that were consequently replicated for the low-autonomy participants were morally positive. So instructions to the 'low autonomy' participants mostly meant to 'do the right thing'. Knowing that things would be morally acceptable could have served as rationalization for potential dissonance in low-autonomy users about the instructions on how to behave in a given decision situation. If the instructions would have more frequently said to do the morally wrong thing, participants in the low-autonomy condition may have felt greater unease about their reduced autonomy, which would technically have resulted in a greater power of experimental manipulation of autonomy. From this perspective, the manipulation power of the "Beergarden" experiment was greater, because the dialogue management that was affected by the manipulation was present and in function throughout most of the exposure time. It made a difference to the technical level of user interaction, whereas in the present "Fable" study, the manipulation was limited to a narrative or content dimension. Moreover, the

manipulation was effective only at four limited points in time during the playing sequence that lasted for 20 to 30 minutes.

From this analysis, we conclude that the present experimental zero-findings do not indicate an insufficiently low sensitivity of the toolkit scales for variation in user experiences. Because the scales turned out to be sensitive with the same number of participants in the “Beergarden” study, we argue that the “Fable” experiment involved a very weak manipulation of the independent variable “user autonomy”. A greater test power (i.e., a larger sample) would possibly have revealed a notable group difference in role adoption / identification, which would then be perfectly in line with theoretical assumptions. Future studies should build on the present design and a) involve more male participants who might be ready to decide for the morally less desirable options and thus increase variance in decision morality across both experimental conditions and b) increase sample size to see whether the observed (weak) tendendy in role adoption / identification data can be stabilized towards an interpretable effect.

2.5 Summative Instrument Assessment: Capabilities and Prospects of the IRIS User Experience Measurement Toolkit

With the studies reported in the previous sections, VUA's empirical research programme for the purpose of development, validation, and performance assessment of the envisioned Measurement Toolkit for user-centered evaluation studies in Interactive Storytelling is complete. Six experimental studies have been conducted that involved five different manifestations of 'Interactive Storytelling' and overall N = 316 participants (the last two studies involved the same 42 individuals, but technically, each study contributes to the overall data base, also in terms of statistical cases).

The purpose of this accumulation of empirical data produced by the same measurement instrument was to critically test whether a theoretically derived set of measures would turn out statistically reliable and conceptually informative across different scenarios of research application. Maximizing variation among the manifestations of Interactive Storytelling was therefore an important goal of the research programme. With systems investigated as diverse as the "Fahrenheit" and "Fable" video games, the foundational (and somewhat historic) IS prototype "Facade", and the more recent IS projects "EmoEmma" and "Virtual Beergarden" (both investigated in cooperation with IRIS partners), the objective of covering a considerable bandwith of systems and media within the landscape of Interactive Storytelling has been achieved. The strategy of employing experimental variations to all investigated systems allowed to generate additional, conceptually relevant outcomes that both help to assess the measurement instrument's strengths and weaknesses and to learn more about the entertainment theory issues involved in user responses to interactive stories. These learning outcomes have so far mostly been addressed based on individual experiments or by comparing results from two of the six studies (Roth et al., in press). Future exploitation of this treasure of empirical data can thus include more aggregate examinations across, for instance, four of the six or even all conducted studies.

Such meta-analytic or combined approaches will primarily serve concept-oriented research purposes, however. The present report's major focus is on methodological quality of the Measurement Instrument and the question whether the developed self-report scales are ready for dissemination among the IS research and development community, as it is foreseen in the WP7 workplan. From this perspective, the synthesized view on how the measurement scales performed across the different studies is overall very positive. The endeavour of deriving empirical measures from an elaborate theoretical framework (D7.1) that 'function well' in terms of statistical reliability, sensitivity to systematic variations of experience-relevant design elements (e.g., interactivity on/off, or actor versus ghost mode), and also in terms of comprehensiveness and practicability has been completed mostly successful. The limitation of "mostly" successful needs to be mentioned, because one of the 13 scales that have been worked on within WP7, the measure of perceived character believability, did not return as satisfying results as the other parts of the toolkit. This is especially true for the more recent studies in which only very short item lists were applied. Given the complex, multidimensional nature of character believability, more conceptual work needs to be done to re-work a better-functioning scale. Some literature on theorizing and measuring character believability is of course available that will help to fix this one remaining problem with the toolkit elements (e.g., Riedl & Young, 2005)

Another important insight is that, as it should be expected from the general social-scientific methods literature, longer scale versions tended to perform better in terms of reliability than shortened versions. For future applications of the toolkit scales, the trade-off between better reliability and shorter, more user-friendly questionnaires will emerge as important decision to

make in defining a research design and protocol; the documentation that will accompany the final measurement toolkit will thus need to direct reseachers' attention towards this issue. By the same token, the final toolkit should include shorter and longer versions of all the measurement scales in order to empower researchers to make individual choices in this reliability/length tradeoff. For specific research issues that focues on pre-defined dimensions of the user experience, it will be necessary, for instance, to generate a maximum reliability for the measures of the concepts that are most crucial. These measures would then be employed in their 'long' version; measures of secondary conceptual relevance, in contrast, could be selected in their 'short version'.

With these considerations, the summary of the empirical work programme of WP7 under the responsibility of VUA is positive. The quality objectives have been met with some minor limitations that need further addressing. Substantial empirical insight has been generated that awaits theoretical interpretation and discussion (e.g., Roth et al., in press; Endrass et al., submitted). From a methodological point of view, the nearly-optimized social-scientific measurement toolkit is now ready to be transformed into a technical measurement toolkit that other research teams can employ for their own purposes in user-centered studies, both formative investigations during system development and outcome-oriented investigations after system development (see also section 4).

3. WP7 Studies as Benchmarking Reference: Overview of Toolkit Results

Following the demonstration that the results produced by application of the measurement toolkit scales are reliable, practicable, and interpretable, the next preparatory step before the release of the toolkit as a convenient research instrument to the IS community is to compile an overview of how participants of VUA's six studies that applied toolkit measures rated the different experiential dimensions across the diverse IS systems and prototypes used in the experiments. This compilation serves as benchmarking reference: If researchers employ the toolkit in their own work, for example, to inspect the strengths and weaknesses of a new IS prototype from the user perspective, they can compare the data obtained from their study with the reference values provided from the different IRIS WP7 studies. Based on necessary prior conceptual considerations on which of the systems investigated so far is more or most similar to one's own prototype, the comparison of future studies' findings with the current reference values will allow understanding where users see positive aspects in the new system and where there is a need (or a chance) to improve.

Therefore, the following table 12 (next page) provides an overview of the mean ratings of those scales that belong to the IRIS measurement toolkit. For each of the six studies, only mean ratings are offered as reference values; these are reported for all experimental conditions of each study. No 'total' values per study are given, because due to the experimental manipulations applied, the total values would mix up user-centered findings from different system versions. A future IS system that other researchers want to investigate with the toolkit may be similar, for example, to the 'ghost mode', but not the 'actor mode' version of the "EmoEmma" system examined in study 4. Total, cross-condition average values for the EmoEmma system would then not be helpful as reference values.

Table 12. Overview of Toolkit scale means across the WP7 studies.

	Study 1: „Fahrenheit“		Study 2: „Facade“		Study 3: „Facade“		Study 4: „EmoEmma“		Study 5: „Beergarden“		Study 6: „Fable“	
User experiences	Inter-activity off	Inter-activity on	Inter-activity off	Inter-activity on	After first exposure	After second exposure	Actor mode	Ghost mode	Round-based dialogue	Continuous dialogue	Low autonomy	High autonomy
A Preconditions												
System usability	3.69	3.11	3.81	3.93	3.66	3.99	4.11	4.22	3.97	3.82	3.76	3.78
Corresp. w. user expectations	3.38	3.63	3.10	3.46	2.89	2.99	3.09	3.35	3.50	3.40	3.86	3.60
Presence	2.62	2.68	2.77	3.27	3.25	3.41	3.26	3.10	---	---	---	---
Character believability	3.48	2.98	3.64	3.84	3.68	3.49	3.12	3.06	2.91	2.58	2.95	3.00
Effectance	2.40	3.23	2.47	3.18	3.00	3.41	2.24	2.88	2.86	3.08	3.40	3.10
B Experiential qualities												
Curiosity	3.43	3.58	3.33	3.49	3.48	3.46	3.59	3.86	3.33	3.78	3.90	3.86
Suspense	3.44	3.33	3.33	3.50	3.53	3.50	3.61	3.49	2.82	3.16	3.80	3.57
Flow	3.00	2.95	2.98	3.00	2.86	3.22	3.09	3.31	4.09*	4.30*	3.54	3.65
Aesthetic pleasantness	2.24	2.00	2.54	2.45	2.45	2.45	2.33	2.44	1.62	1.60	1.48	1.75
Pride	---	---	---	---	2.38	2.57	2.21	3.01	---	---	---	---
Autonomy	---	---	---	---	---	---	2.17	2.47	2.17	2.68	3.35	3.02
Enjoyment	2.80	2.94	2.54	2.86	3.15	3.20	3.68	3.69	3.07	3.45	3.90	3.79
C Specific experience Measures												
Affect: positive	4.51	4.60	4.31	5.00	2.71	2.76	3.00	3.17	2.38	2.20	2.40	2.62
negative	2.91	2.59	4.06	3.05	2.77	2.60	1.97	1.88	1.32	1.46	1.38	1.45
Role adoption	2.67	2.71	2.88	3.24	2.97	2.88	2.76	2.63	1.83	2.07	2.35	2.83

*single item due to low scale reliability

The overview table of course also allows comparing the findings from different studies directly with each other. Much conceptual insight may be generated from analyzing this overview table in terms of user responses to different (versions of) IS systems. However, the theory- and design-related implications are not addressed here beyond the discussions of the individual studies (see section 2). Clearly, the empirical treasure of which the table 12 is a summary is highly interesting to inform the debates on interactive storytelling and entertainment computing at large; the same is true for social-scientific discussions on media entertainment and the psychology of video gaming as well. But for the present purpose of providing reference data for future applications of the toolkit, this overview offers a compact perspective on how to interpret newly generated scale means for other systems and prototypes. Most researchers will also be adviced to read into specific descriptions of WP7 studies and our interpretation of results, but the overview table is an important element to provide orientation about what to expect when utilizing the measurement toolkit in one's own research.

4. Outlook

After 30 months of theoretical and empirical work, WP7 has brought about a rich theoretical framework on user experiences to interactive storytelling as well as a measurement tool that allows exploring user reactions to IS prototypes and media reliably and validly, and with a high degree of practicability. Six experimental investigations provide equally rich empirical knowledge on how the scales perform and what can be expected from diverse types of Interactive Storytelling. The social-scientific approach to user-centered research in Interactive Storytelling that was envisioned for IRIS' WP7 has made significant progress and met the objectives set out up to this point of the network runtime.

The remaining IRIS time will thus be dedicated to two major remaining tasks. One is to transform the conceptual and empirical materials that form the measurement toolkit (scale items, manual of application, documentation of reference studies) into a software solution that researchers in the IS community can easily apply within their own user-oriented studies. The idea is that this software solution will render examinations of user reactions both during system development (formative research) and after project completion (evaluation research) more accessible for research teams who are not strongly connected to social scientific research units and/or not well-equipped to conduct social research by themselves. At the same time, the standardized toolkit software will allow to generate an interesting empirical knowledge base that can easily grow beyond the present six-study data archive into a much larger set of data which then would allow very promising new ways of collaborative analysis and reflection (e.g., in special workshops at future ICIDS or entertainment computing conferences). For this purpose, VUA is working with a programmer to develop a software solution that supports research teams with study planning (e.g., which scales from the toolkit do I need? For which experiential dimensions should I use long-version scales?), data collection (i.e., by enabling computer-supported self-administered interviews), and data analysis (i.e. computation of scale reliabilities and experimental group means in the very same way as data have been presented in the IRIS reports and publications). This software solution will thus offer convenient support to research teams and at the same time serve the beneficial standardization in user-centered research on Interactive Storytelling. Introduction workshops will be held at relevant conferences (one workshop has recently been accepted for ICEC 2011, and another one is being submitted to ICIDS 2011) in order to present the toolkit and motivate colleagues within and outside of the IRIS network to try out the toolkit in their own work.

Second, the gallery of results from six experimental studies with different scopes, but a uniting topic and a shared theoretical framework will be exploited in terms of conceptual reflections about the entertainment experience in Interactive Storytelling. First comparative considerations from two of the six studies will be presented at ICEC 2011 (Roth et al., in press), and additional analyses and manuscripts are foreseen for the final months of the IRIS runtime. This way, WP7 work will not only result in methodological support to the IS community, but also make relevant contributions to the theoretical understanding of user-related issues in Interactive Storytelling. By circulating the conclusions from the six IRIS experiments among the IS community, a closer connection between traditional applied design-focused research on IS on the one hand and media psychology as well as social-scientific entertainment research on the other hand will become a reality.

5. References

Cavazza, M., Lugrin, J.-L., Pizzi, D. and Charles, F., (2007). Madame Bovary on the Holodeck: Immersive interactive storytelling. ACM Multimedia 2007, pp. 651-660

Csikszentmihalyi, M.: Flow: The psychology of optimal experience. New York: Harper Row. (1990)

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)

Kline, P.: A handbook of test construction. Introduction to psychometric design. Methuen, London (1986)

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

Klimmt, C., Vorderer, P. & Ritterfeld, U.: Interactivity and generalizability: New media, new challenges. Communication Methods and Measures, 1 (3), 169-179 (2007)

Klimmt, C., Roth, C., Vermeulen, I., Vorderer, P. & Roth, F. S.: Forecasting the experience of future entertainment technology: "Interactive Storytelling" and media enjoyment. Games and Culture: A Journal of Interactive Media (in press)

Roth, C., Klimmt, C., Vermeulen, I. & Vorderer, P.: The experience of Interactive Stories: Comparing „Fahrenheit" with „Facade". In J. Ancaleto & N. Graham (eds.): Entertainment Computing – Proceedings of the 10th International Conference on Entertainment Computing, ICEC 2011. Springer, Berlin (in press)

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)

Ryan, R. M., Rigby, C.S. & Przybylski, A.: The motivational pull of video games: A self-determination theory approach. Motivation and Emotion, 30, 347-363. (2006)

Tamborini, R., Bowman, N. D., Eden, A., Grizzard, M. & Organ, A.: Defining media enjoyment as the satisfaction of intrinsic needs. Journal of Communication, 60 (4), 758-777 (2010)