PQS 2016 5th ISCA/DEGA Workshop on Perceptual Quality of Systems
29-31 Aug 2016, Berlin, Germany

Longitudinal User Experience of a Mobile Service

Stefan Schaffer[1], *Rene Kelpin*[2], *Norbert Reithinger*[1]

[1]German Research Center for Artificial Intelligence, Intelligent User Interfaces, Germany
[2]German Aerospace Center, Institute of Transport Research, Germany
{stefan.schaffer, norbert.reithinger}@dfki.de, rene.kelpin@dlr.de

Abstract

In this paper we examine influencing factors and changes of perceived longitudinal user experience. An implemented mobility app for Berlin was tested in a three month field trial. The app offers unique features for bicyclists and integrates a gamification concept. The study consisted of three game phases. Each phase was completed by a demographic and a user experience questionnaire. Possible differences resulting from differing gamification incentives between acquired participants and voluntary participants could not be found. Further, general changes in subjective long-term user experience could not be detected. The results reveal several significant effects of age, gender, and specific user groups on pragmatic quality, attractiveness, and hedonic qualities.

Index Terms: user experience, long-term studies, mobile service

1. Introduction

The concept of user experience (UX) has undergone a rapid development in the recent years. As part of a human-centered development of technology UX is a central issue that must be observed in order to gain a competitive advantage in the market, and to utilize the positive effects that technology offers for everyone today [1]. Previous research identified and analyzed various aspects of UX, as e.g. aesthetics, emotions and other experiential aspects [2]. Especially for long-term usage certain aspects may be more relevant for specific user groups than they appear at first. For example, experiences that can be induced by a technology in the first few days may differ from the experiences perceived by the user after a longer use.

Within the EU project STREETLIFE we developed a mobility application that supports its users in finding CO_2-saving ways in the city of Berlin [3]. The aim of the project is to demonstrate that unique features for cyclists, such as "avoiding accident hotspots" and special gamification elements lead to an increased use of the bicycle and thus result in CO_2-savings [4, 5]. In order to examine the impact of the proposed solution it is necessary to conduct a longitudinal study assessing the mobility behavior of the system users. This implies, however, that compared to the size of Berlin a sufficiently large number of users is needed participating in such a study. In addition, the users should continue and not abort their participation in the study.

For Berlin already a number of apps supporting users during daily mobility exist. These apps enable their users to efficiently utilize the cities infrastructure, while also ensuring good UX. It must be assumed that highlighting unique features of the STREETLIFE app alone, will not be sufficient to get the users to continuously using the app. To gain and keep a sufficiently large number of users it is necessary to provide the users with acceptable UX (compared to the other Berlin apps).

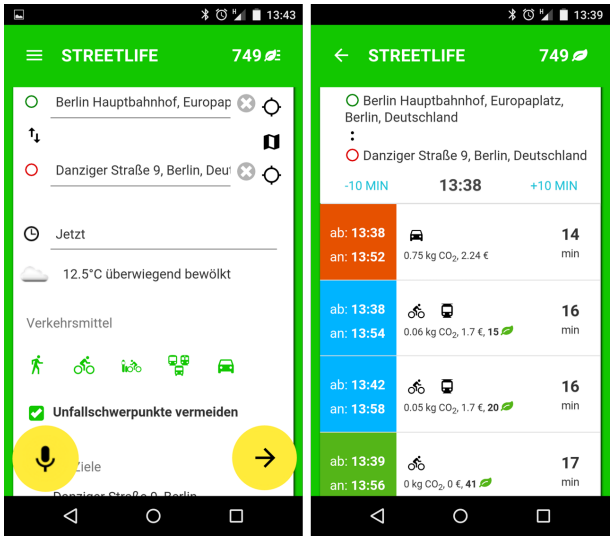

Figure 1: *Left: routing view. Right: routing proposal view.*

We therefore conducted a long-term study to gather relevant data for the assessment of different impact categories of the project. Along with CO_2-savings an important interest of our research are long-term evaluation methodologies in HCI. The ongoing study therefore examines changes of UX aspects that may occur during long-term use. The aim of this paper is to conduct an explorative analysis examining various UX aspects at different times of measurement. Concretely we face the following exploratory research questions:

1. What trends can be observed for UX ratings of user groups differing in gamification incentives?
2. To what extent can questionnaires detect changes in subjective longterm UX?
3. What demographic factors determine the UX?

2. Longitudinal Study

2.1. The STREETLIFE Mobility App

The main purpose of the mobility app is to efficiently support users stating route requests and to help selecting appropriate itineraries. Figure 1 provides an overview of the most important app views. In the view on the left users can input start, destination, preferred transportation modes, travel time and the optional "avoid accident hotspot" feature for bicyclists. Fur-

http://dx.doi.org/10.21437/PQS.2016-28

ther the actual number of points (leaves) is displayed in the upper right corner of the screen. The routing proposals are presented in a list containing information about duration, CO_2 emission, monetary costs, traffic modes, and points to get for the itineraries (Figure 1, right). The gamification concept foresees 10 leaves for each cycled kilometer. The STREETLIFE routing is also able to find so called inter-modal routes combining e.g. bicycle with public transportation. The colors at the left side of each of the routing proposals indicate the "eco friendliness" of the itinerary. The details of the itineraries can be viewed in textual form and the routes can be displayed on a virtual map. By pressing a navigation button a "Companion Mode" is activated. During the companion mode the current position and the selected route are displayed on a map. The arrival at a destination is automatically detected by the app. If a trip ends the users get their points for bicycling. As part of the gamification concept the "top-ten" users can be viewed in the app and virtual trees can be planted on a virtual map of Berlin. The virtual trees are issued for every 500 collected leaves and can be seen by all users. Users participating in the gamifiaction have to choose a nickname. Further they can state an email address to take part in the evaluation of the project. The App has been realized using an HTML5-based framework [6], and is available free of charge at the Google Play Store[1].

2.2. Participants

With regard to the acquisition of test persons the participants of the longitudinal study can be divided into two groups. The first group was acquired by a test person agency. The participants of the second group downloaded the app voluntary from the app store. In order to gain a sufficiently large number of participants several public relation activities promoting the app and the study were undertaken. Participants acquired by the agency had to state their email address in the app. The other app users could decide on their own if they wanted to provide their email address or not. Only participants who provided a valid email address could receive the email invitation to the user study and take part in the evaluation. Some of the acquired participants discontinued their participation during the execution of the study. Therefore the number of acquired participants decreased sightly. The number of voluntary participants, however, rather increased, as more and more users installed and used the app during the evaluation period. Table 2 summarizes the basic statistics of the participants.

2.3. Study Design

The acquired test persons had to be physically present at an introductory meeting where they were informed about the project, the app, and the conditions of participation. In contrast the voluntary participants obtained their information only from the app descriptions in the app store, and the descriptions within the app, as well as from the public relation activities.

As part of the gamification concept the study was conducted in three phases (lasting from the beginning of March until end of May 2016). Each of the game phases lasted one month. At the end of a game phase prices were raffled among the top-ten players. The test persons acquired by the agency also received an 15 Euro Amazon voucher after the end of the last game period. As a further incentive they had the chance to win one of two tablets after the last game phase.

After each game phase a demographic questionnaire and the AttrakDiff UX questionaire [7] were asked in the format of an online survey using "lamapoll"[2]. By means of the demographic data the participants were clustered into roughly equal sized groups using the following categories:

- Age: younger participants (18-29 years) / older participants (30+ years)
- Gender: male / female
- Bicycle usage: frequent (at least once per week) / rare (less than once per week)
- Public transportation usage: frequent (at least once per week) / rare (less than once per week)
- Car usage: frequent (at least once per month) / rare (less than once per month)[3]

The AttrakDiff was used as a standardized questionnaire in order to gather information about the following aspects of perceived UX:

- Attractiveness (ATT): e.g. how good, bad, beautiful or ugly a product is experienced
- Hedonic quality - identity (HQ-I): addresses the need of self expression and being perceived by others in a certain way
- Hedonic quality - stimulation (HQ-S): extent to which the system simulates the need for personal development (e.g. new skills and knowledge)
- Pragmatic quality (PQ): e.g. ease of use, usefulness and usability

In the next section the three time of measurement are referred to as:

- T1: measurement end of March
- T2: measurement end of April
- T3: measurement end of May

Table 1: *Basic statistics of the participants, including the total number of participants N, the number of female participants, the mean age M_{age}, and the standard deviation of the age SD_{age}. The survey included three times of measurement: T1, T2, and T3.*

	T1			
Participants	N	female	M_{age}	SD_{age}
acquired	34	20	33,30	11,10
voluntary	8	3	39,29	10,50
	T2			
Participants	N	female	M_{age}	SD_{age}
acquired	31	19	33,52	11,06
voluntary	16	6	44,93	12,58
	T3			
Participants	N	female	M_{age}	SD_{age}
acquired	32	18	32,94	11,07
voluntary	15	4	39,79	18,43

[1] https://play.google.com/store/apps/details?id=de.dfki.iui.mmir.streetlife

[2] https://www.lamapoll.de

[3] In order to arrive at two roughly equal sized groups the frequency of car usage comprises a larger period of time for the group with frequent car usage.

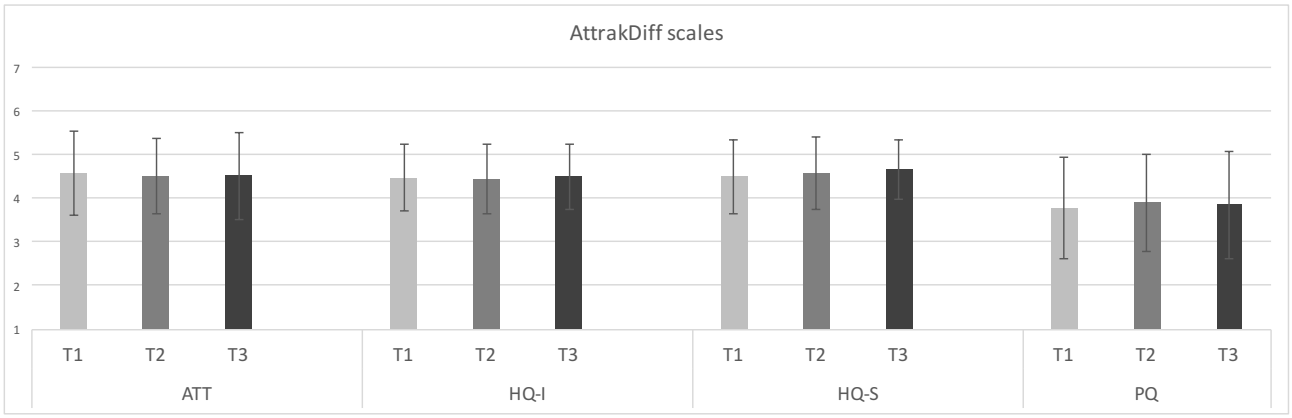

Figure 2: Results of the AttrakDiff scales for the three times of measurement T1, T2, and T3. Error bars illustrate standard deviations (SD). Bars indicate mean values (M).

3. Results

3.1. Participants and times of measurement

In order to examine possible differences between acquired participants and voluntary participants T-tests were calculated for all AttrakDiff scales at the three times of measurement. No significant differences could be detected. A one-way ANOVA with repeated measures was conducted to investigate possible differences between the three times of measurement for all AttrakDiff scales. Again no significant differences could be revealed. Figure 2 illustrates the marginal differences between the means and standard deviations of the single AttrakDiff scales at the different times of measurement.

3.2. Demographic Factors

3.2.1. Age

Most evident effects can be reported for the two age categories. At T1 23 subjects were clustered in the category of older participants and 19 subjects were clustered into the category of younger participants. For PQ at T1 within the group with older participants higher values (M = 4,17, SD = 1,04) than in the group with younger participants (M = 3,29, SD = 1,16) could be observed. The difference was significant (t(40) = -2,604, p = 0,013).

At T2 the group with older participants assessed attractiveness, hedonic quality - identity, and pragmatic quality significantly better than younger users. Cluster size, means, standard deviations and the results of the T-tests are reported in Table 2.

Table 2: *Cluster size, means, standard deviations and the results of the T-tests for age categories at T2.*

		N	M	SD	t(45)	p
ATT	young	18	4,05	0,87	-3,438	0,001
	old	29	4,85	0,72		
HQ-I	young	18	4,05	0,82	-3,050	0,004
	old	29	4,71	0,67		
PQ	young	18	3,36	1,07	-3,134	0,003
	old	29	4,32	0,98		

At T3 27 subjects were clustered in the category of older participants and 20 subjects were clustered into the category of younger participants. For attractiveness at T3 within the group with older participants higher values (M = 4,80, SD = 1,12) than in the group with younger participants (M = 3,94, SD = 0,89) could be observed. The difference was significant (t(45) = -2,869, p = 0,006). Further for pragmatic quality at T3 within the group with older participants higher values (M = 4,22, SD = 1,28) than in the group with younger participants (M = 3,18, SD = 1,09) could be observed. The difference was significant (t(45) = -2,921, p = 0,005).

3.2.2. Public Transportation Usage

At T2 26 subjects were clustered in the category of frequent public transportation users and 18 subjects were clustered into the category of rare public transportation users. For pragmatic quality at T2 within the group of rare public transportation users higher values (M = 4,42, SD = 1,01) than in the group of frequent public transportation users (M = 3,57, SD = 1,06) could be observed. The difference was significant (t(45) = -2,812, p = 0,007).

At T3 29 subjects were clustered in the category of frequent public transportation users and 21 subjects were clustered into the category of rare public transportation users. For pragmatic quality at T3 within the group of rare public transportation users higher values (M = 4,31, SD = 1,39) than in the group of frequent public transportation users (M = 3,44, SD = 1,14) could be observed. The difference was significant (t(45) = -2,325, p = 0,025).

3.2.3. Car Usage

At T2 27 subjects were clustered in the category of frequent car users and 20 subjects were clustered into the category of rare car users. For pragmatic quality at T2 within the group of frequent car users higher values (M = 4,23, SD = 1,02) than in the group of rare car users (M = 3,57, SD = 1,15) could be observed. The difference was significant (t(45) = 2,069, p = 0,044).

At T3 also 27 subjects were clustered in the category of frequent car users and 20 subjects were clustered into the category of rare car users. For pragmatic quality at T3 again within the group of frequent car users higher values (M = 4,14, SD = 1,09) than in the group of rare car users (M = 3,28, SD = 1,42) could be observed. The difference was significant (t(45) = 2,364, p = 0,022).

3.2.4. Gender

At T2 24 female and 23 male participants were clustered into two groups. For hedonic quality - stimulation at T2 within the male participants higher values (M = 4,90, SD = 0,79) than within the female participants (M = 4,32, SD = 0,77) could be observed. The difference was significant (t(45) = 2,549, p = 0,014).

4. Discussion

Contrary to our assumption for acquired and voluntary participants no significant differences on the AttrakDiff scales could be found. Although the acquired test persons received more incentives than the voluntary test persons they did not rate the perceived quality of the system better. An explanation for this result could be that the voluntary test persons did not know about the acquired group during the evaluation. This result gives an indication that the perceived user experience is not necessarily correlated with the amount of gamification incentives. Further, as they were not acquired, the voluntary participants independently decided to install and check out the mobility app. This special motivation may also cause a tendency for a better subjective assessment if the perceived experiences are positive. Positive experiences could also compensate the negative effect that may arise from too low incentives.

No significant differences in user experience could be detected between the single times of measurement. From a system evaluation point of view this result could be seen as a validation of a stable user experience; a desirable feature for consumer products. However, one could also interpret the unchanged AttrakDiff scales as an indication that the gamification part does not especially engage the system users. If this is desired other strategies than the implemented game concept have to be conceived.

Interestingly older users rated pragmatic quality significantly better than younger users over all times of measurement. The mean assessment of PQ of older users is over 4,0 in all cases. This could imply that the overall usability of the app can be categorized in the upper middle field. However, younger participants assessed PQ below 3,5 in all cases. As younger people often have an affinity to try out the newest and trendiest applications they often also have a better feeling for modern interaction an UX designs. If this is the case here, this could mean that the app does not implement the actual interaction and UX design trends in a good enough manner.

Further at T2 and T3 older users rated the attractiveness significantly better than younger users. It has to mentioned that this results mainly from a relatively stable rating of the older users, and consistently decreasing ratings of the younger users at every time of measurement (with a younger user mean attractiveness of 4,27 at T1). From a system evaluation point of view this development of the younger users ATT scale assessment can be interpreted as alarming. The reason for this decrease is not revealed by the evaluation. One possible source could be the gamification concept. It has been developed in a way that only the top-ten players can receive real awards. Therefore an average player experiencing that it is harldy possible to get into the top-ten with average gaming performance could find the app less attractive.

At T2 older users rated HQ-I significantly better than younger users. A similar tendency as for the ATT scale can in his case not be observed. HQ-I shows variation for younger and older users at all times of measurement. It would be interesting to re-examine the development of this scale after a longer time period.

Interestingly the group of rare public transportation users significantly better assessed pragmatic quality at T2 and T3. One reason for this result could be that the user group got used to the new inter-modal routing concept. The combination of other traffic modes with public transportation may constitute a reasonable alternative for this user group compared to an exclusive public transportation usage.

A further interesting result is that frequent car users also significantly better assess pragmatic quality at T2 and T3. An attempt at an explanation could be that the app shows the CO_2 emission for each itinerary. Perhaps the frequent car users positively notice the potential for CO_2 savings that can be derived by using the app.

5. Conclusion

We examined three research questions in the field of perceived longitudinal user experience testing a mobility app for Berlin in a three month field trial. The app especially supports bicyclists and integrates a gamification concept. After eech of three game phases demographic and UX data was subjectively assessed. Research question 1 investigated if trends can be observed for UX ratings of user groups that differ in gamification incentives. Possible differences resulting from differing gamification incentives between acquired participants and voluntary participants could not be found.

In the second research question it was observed to what extent questionnaires can detect changes in subjective longterm UX. None of such changes could be detected by the conducted field trial.

The third research question examined demographic factors that determine UX. Significant effects of age on pragmatic quality, attractiveness, and hedonic quality - identification, of gender on hedonic quality - stimulation, and of other specific user groups on pragmatic quality could be found.

The STREETLIFE app will be further operated after the conducted field trial. This enables further measures to investigate longterm UX.

6. Acknowledgements

This work is part of the STREETLIFE (Steering towards Green and Perceptive Mobility of the Future) project co-funded under the 7th RTD Framework Programme, FP7-SMARTCITIES-2013 – grant agreement 608991.

7. References

[1] M. Minge, "Dynamische Aspekte des Nutzungserlebens der Interaktion mit technischen Systemen," Ph.D. dissertation, TU Berlin, 2011.

[2] M. Hassenzahl and N. Tractinsky, "User experience-a research agenda," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, mar 2006.

[3] S. Schaffer, A. Ruß, and N. Reithinger, "User Guided Speech Technology Integration for a Mobility Application," in *accepted at DSLI workshop at CHI '16*, San Jose, CA, USA, 2016.

[4] S. Schaffer and N. Reithinger, "Intermodal personalized travel assistance and routing interface," in *Mensch und Computer 2014*, A. Butz, M. Koch, and J. Schlichter, Eds. De Gruyter Oldenbourg, 2014, pp. 343–346.

[5] A. Nurminen, K. Sirvio, S. Schaffer, A. Marconi, and G. Valetto, "End-user applications techniques and tools (intermediary)," Tech. Rep., 2015.

[6] A. Ruß, "MMIR Framework: Multimodal Mobile Interaction and Rendering." in *GI-Jahrestagung 2013*, 2013.

[7] M. Hassenzahl, M. Burmester, and F. Koller, "Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität," *Mensch & Computer 2003: Interaktion in Bewegung*, 2003.