Qualitative Evaluation - Hypothesis, Test cases, and Metrics

Hypothesis

According to Nielsen, usability can be divided into five elements, i.e., learnability, efficiency, memorability, errors, and satisfaction[(1)]. And the ISO 9241-11 standard suggests that usability metrics should include effectiveness, efficiency, and satisfaction[(2)]. To combine these two approaches, we developed hypotheses for evaluating the "TU-tatsy" application in six areas: Learnability, Effectiveness, Efficiency, Memorability, Errors, and Subjective Satisfaction.

Goals	Hypothesis
1. Learnability	**1.1** More than one way is used to cancel dishes.
2. Effectiveness	**2.1** Average completion rate of >70% for all tasks completely performed by users
	2.2 At least 80% of the test persons complete the task of book dishes.
	2.3 At least 80% of the test persons complete the task of checking the location of the counter where the dish is located and finding out the queue at this counter.
3. Efficiency	**3.1** The average time for a user to book a dish for the first time is less than 20s.
	3.2 The average time to cancel a scheduled dish is at most the 20s.
4. Memorability	**4.1** At least 70% of users can remember the three first-level navigation of the product after testing.
5. Errors	**5.1** The average number of errors is at most 6 in fulfilling all tasks.
	5.2 At most 40% of the test persons will be confused or make mistakes with the counter map.
6. Subjective satisfaction	**6.1** The prototype achieves at least a rating of 6.5 out of 10 on average.
	6.2 SUS-Score exceeds a value of 70.
	6.3 The average Task Level Satisfaction does exceed a value of 5.

Test cases

In this step, the test cases are generated based on the UML use cases created in iteration-2. The purpose of these tasks is to accept or reject the hypotheses in the last subsection.

Task	Instruction
1	Please follow the link and start the app.
2	Please view the mensa occupancy.
3	Please view today's meal schedule.
4	Please check the location of the counter where the dish "Spaghetti with tomato sauce and spinach" is located and find out the queue length at this counter.
5	Please view more information on “Spaghetti with tomato sauce and spinach”.
6	Please view the most recent review of this dish.
7	Please check the ingredients, the rating and the allergens for “Spaghetti with

	tomato sauce and spinach”.
8	Please book “Spaghetti with tomato sauce and spinach”.
9	Please cancel your reservation for "Spaghetti with tomato sauce and spinach".
10	Please name as many system features (functionalities) of the app that you remember. You have 30 seconds to do this.

Metrics

- Objective measures

Goals	Hypothesis	Objective measures
1. Learnability	**1.1** More than one way is used to cancel dishes.	The number of different commands to fulfill a specific task.
2. Effectiveness	**2.1** Average completion rate of >70% for all tasks completely performed by users	$\frac{number\ of\ task\ \ which\ completed\ successfully}{10} \times 100\%$
	2.2 At least 80% of the test persons complete the task of booking dishes.	$\frac{number\ of\ people\ who\ completed\ the\ task}{number\ of\ people\ who\ did\ \ not\ complete} \times 100\%$
	2.3 At least 80% of the test persons complete the task of checking the location of the counter where the dish is located and find out the queue at this counter.	$\frac{number\ of\ people\ who\ completed\ the\ task\ 5}{total\ number\ of\ test\ persons} \times 100\%$
3. Efficiency	**3.1** The average time for a user to book a dish for the first time is less than 20s.	The time users take to complete task 8.
	3.2 The average time to cancel a scheduled dish does not exceed 20s.	The time users take to complete task 9.
4. Memorability	**4.1** At least 70% of users can remember the three first-level navigation of the product after testing.	$\frac{number\ of\ \ people\ \ who\ \ remembered\ \ navigation\ levels}{total\ number\ of\ test\ persons}$ $\times 100\%$
5. Errors	**5.1** The average number of errors does not exceed 6 in fulfilling all tasks. (wrong tasks executed, wrong clicks, too many clicks)	The number of errors for all tasks.
	5.2 At most 40% of the test persons will be confused or make mistakes with the counter map.	$\frac{number\ of\ \ people\ \ who\ \ make\ error\ by\ task\ 4}{total\ number\ of\ test\ persons} \times 100\%$
6. Subjective satisfaction	**6.1** The prototype achieves at least a rating of 6.5 out of 10 on average.	Rating number
	6.2 SUS-Score exceeds a value of 70.	SUS-Score
	6.3 The average Task Level Satisfaction does exceed a value	SEQ-Score

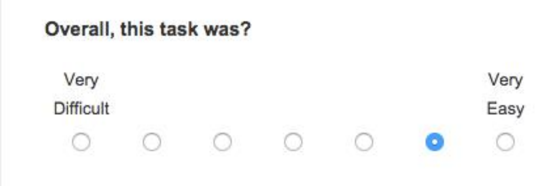
	of 5[5].	

- Subjective measures

1. Question: How would you rate the prototype overall on a scale of 1-10? → to get an overall rating.
2. Task-level satisfaction

   After each task is completed, the test persons are asked to rate the Single Ease Question (SEQ)[3], which is often recommended for task-level satisfaction for its ease and correlation with other usability metrics.

   It consists of just one question after a task:



Overall, this task was?

Figure 1: The Single Ease Question (SEQ).

3. Test-level satisfaction

   At the end of the session the test persons are asked to rate the 10 statements of the System Usability Scale (SUS)[4], the overall SUS score is used to assess usability[1].

   The 10 items with one of five responses that range from Strongly Agree to Strongly disagree can be viewed below:

   (1) I think that I would like to use this system frequently.
   (2) I found the system unnecessarily complex.
   (3) I thought the system was easy to use.
   (4) I think that I would need the support of a technical person to be able to use this system.
   (5) I found the various functions in this system were well integrated.
   (6) I thought there was too much inconsistency in this system.
   (7) I would imagine that most people would learn to use this system very quickly.
   (8) I found the system very cumbersome to use.
   (9) I felt very confident using the system.
   (10) I needed to learn a lot of things before I could get going with this system.

Likert-Scale

Strongly disagree				Strongly agree
1	2	3	4	5

Experimental Setup

Location/Tools: Online via Zoom, Chrome Browser.
Test persons: A total of 10 people, including tum students and faculty members, with a balanced gender ratio.
Data Acquisition:

Data recording with Computer Microphone and Camera (screen and voice recording), Timer, and Counting events (e.g. errors) by hand.

Procedure

1. Test persons are welcomed, briefly self-introduction, briefly introduced to the topic of the testing.
2. Link to the app will be shared; Test persons are asked to share his/her screen during the experiment.
3. Introduce the tasks and considerations of this test.
4. Permission for recording is obtained, and screen and voice are recorded.
5. Test persons conduct tasks one by one using the think aloud method, the experimenter will not be allowed to help.
6. Test persons fill out the relevant questionnaire and answer the questions.
7. Organizing data, compiling all the data in an Excel File and analyzing it with R or SPSS, and compared with the hypothesis.

References

[1] ” Usability 101: Introduction to Usability”, https://www.nngroup.com/articles/usability-101-introduction-to-usability/, Retrieved on 12.01.2023

[2] ISO 9241-11 (2018) „Usability: Definitions and concepts”

[3]”10 Things To Know About The Single Ease Question (SEQ)”, https://measuringu.com/seq10/, Retrieved on 12.01.2023

[4] “System Usability Scale (SUS)”, https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html, Retrieved on 12.01.2023

[5] “Usability Metrics – A Guide To Quantify The Usability Of Any System”, https://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/, access on 09.01.202