Hypothesis, test cases and metrics

Hypothesis

The goal of a summative evaluation, assessing overall quality of an interface, is a very abstract goal, which is why it must be broken down into manageable and, above all, measurable components. According to Nielsen, usability can be divided into five elements, namely learnability, efficiency, memorability, errors, and satisfaction.[1] The ISO 9241-11 standard, on the other hand, recommends that usability metrics should include effectiveness, efficiency and satisfaction.[2]

In order to combine these two approaches, hypotheses were built for the evaluation of the app "Craftive" on the six components Learnability, Effectiveness, Efficiency, Memorability, Errors and Subjective Satisfaction.

Goals	Hypothesis
1. Learnability	**1.1** More than one way is used to access a user´s profile by a test person.
	1.2 The user can complete the creation of a second request faster than the creation of the first request.
2. Effectiveness	**2.1** At least 78% of the test persons manage the task of checking the rating of a user´s profile.
	2.2 At least 78% of the test persons manage the task of deleting a request.
	2.3 At least 78% of the test persons manage the task of agreeing on a deal (paying for one´s help offer).
	2.4 At least 78% of the test persons manage the task of viewing the amount of their own liked projects (on their profile page).
3. Efficiency	**3.1** The average time of creating a new request (high level performance task) does not exceed 100 seconds.
	3.2 The average time of getting to and liking a specific project (medium level performance task) does not exceed 40 seconds. (multiple possible options to fulfill this task)
	3.3 The amount of time the user reads the introduction pages is longer than 10 seconds on average.
4. Memorability	**4.1** 60% of the test persons can remember the steps to fulfill the task of texting a possible helper during a debriefing after the test.
	4.2 On average all users can remember at least 4 system features within 30 seconds during a debriefing after the test.
5. Errors	**5.1** At most 70% of the test persons will make an error while commenting a specific project on the first try.
	5.2 At most 70% of the test persons will make an error while texting an interested helper a message.
	5.3 The number of user errors in fulfilling all tasks will not exceed 20 on average.
	5.4 The time spent recovering from errors will not exceed 30 seconds on average.
	5.5 The amount of “dead” time when the user is not interacting with the system (thinking—time delays) for all tasks does not exceed 30 seconds on average.

[1] Nielsen, Jakob (1993). Usability Engineering || What Is Usability?., (), 23–48. doi:10.1016/b978-0-08-052029-2.50005-x (Retrieved on 14.01.2022)

[2] https://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/ (Retrieved on 14.01.2022)

6. Subjective satisfaction	**6.1** The prototype achieves at least a rating of 6 out of 10 on average.
	6.2 The number of times the user expresses clear frustration does not exceed 3 times on average.

Test cases

In the next step, test cases (tasks) were derived based on the use cases created in the previous iterations. The tasks should also help the development team to prove or reject the previously established hypotheses and are therefore closely linked to them.

Task	Instruction
Task 1	Please follow the link and start the app (→ introduction page)
Task 2	Please release a request with the following dates: • Title: Paint walls of my kitchen • Keyword: Painting works • Post Code: 80995 • Request deadline: 02.02.2022 • Description: Can someone help me painting the walls in my kitchen? • Photo: you don´t have to select one
Task 3	Please delete one of the already existing requests
Task 4	Please select one person to help you with the problem of Audi A6´s front brakes and fulfill the steps to enter into the deal
Task 5	Please send a message to an interested helper saying “Thanks for your help”
Task 6	Please check the exploring page
Task 7	Access/check another user´s profile
Task 8	Please check the rating of a user´s profile
Task 9	Please choose another user´s project and give it a like
Task 10	Please comment another user´s project saying “Great work”
Task 11	Please have a look at the amount of projects you liked with your own profile
Task 12	Please release a request again with the following dates: • Title: Tree felling • Keyword: Tree felling • Post Code: 80995 • Request deadline: 17.02.2022 • Description: Can someone help me felling a tree in my garden? • Photo: you don´t have to select one
Task 13	Please tell me the steps you go through to fulfill the task of texting a possible helper from the starting point of opening the app

Task 14	Please name as many system features (functionalities) of the app that you remember. You have 30 seconds to do this.

Metrics

Objective measures

The definition of objective measures was mainly based on the typical quantifiable usability measurements proposed by Nielsen.[1]

Goals	Hypothesis	Objective measures
1. Learnability	**1.1** More than one way is used to access a user´s profile by a test person.	The number of different commands/features to fulfill a specific task that were utilized by the user.
	1.2 The user can complete the creation of a second request faster than the creation of the first request.	The ratio between the number of cases in which the statement is true and the number of cases in which it is false. (No =0, Yes = 1)
2. Effectiveness	**2.1** At least 78% of the test persons manage the task of checking the rating of a user´s profile.	The ratio between the number of tasks completed successfully and the total number of tasks undertaken.
	2.2 At least 78% of the test persons manage the task of deleting a request.	The ratio between the number of tasks completed successfully and the total number of tasks undertaken.
	2.3 At least 78% of the test persons manage the task of agreeing on a deal (paying for one´s help offer).	The ratio between the number of tasks completed successfully and the total number of tasks undertaken.
	2.4 At least 78% of the test persons manage the task of viewing the amount of their own liked projects (on their profile page).	The ratio between the number of tasks completed successfully and the total number of tasks undertaken.
3. Efficiency	**3.1** The average time of creating a new request (high level performance task) does not exceed 100 seconds.	The time users take to complete a specific task.
	3.2 The average time of getting to and liking a specific project (medium level performance task) does not exceed 40 seconds. (multiple options to fulfill this task)	The time users take to complete a specific task.
	3.3 The amount of time the user reads the introduction pages is longer than 10 seconds on average.	Time spent using the help system.
4. Memorability	**4.1** 60% of the test persons can remember the steps to fulfill the task of texting a possible helper.	Ratio between the number of users who can remember the steps for fulfilling a specific task and the number of users who cannot.
	4.2 On average all users can remember at least 4 system	The number of system features the user can remember within a specific time period during a debriefing after the test.

	features within 30 seconds during a debriefing after the test.	
5. Errors	**5.1** At most 70% of the test persons will make an error while commenting a specific project on the first try.	Ratio between the users who will make an error fulfilling a task and the users who will not.
	5.2 At most 70% of the test persons will make an error while texting an interested helper a message.	Ratio between the users who will make an error fulfilling a task and the users who will not.
	5.3 The number of user errors in fulfilling all tasks will not exceed 20 on average.	Number of errors for all tasks.
	5.4 The time spent recovering from errors will not exceed 30 seconds on average.	Time spent recovering from errors.
	5.5 The amount of "dead" time when the user is not interacting with the system (thinking—time delays) for all tasks does not exceed 30 seconds on average.	Amount of "dead" time (thinking-time delays)
6. Subjective satisfaction	**6.1** The prototype achieves at least a rating of 6 out of 10 on average.	Rating number
	6.2 The number of times the user expresses clear frustration does not exceed 3 times on average.	Number of times the user expresses clear frustration

Subjective measures

For the subjective measures, the first question asked was about the overall rating of the prototype ("How would you rate the prototype overall on a scale of 1-10?").

This is followed by a subjective measurement in which the test persons are asked to rate the statements of the SUS (System Usability Scale).

The SUS by Brooke (1996) was chosen because it is a simple usability scale that gives a global view of subjective assessments of usability with the help of only 10 questions.[3] The fact that it is a short questionnaire was important because the objective measurement part is already quite extensive, and the test subjects should not be overloaded.

Furthermore, the SUS has already been used for a variety of research projects and industrial evaluations. The statements of the SUS cover a range of aspects of system usability, such as the need of support, training, and complexity, but at the same time, the questions are not too in-depth.[3] This is another advantage of the SUS questionnaire in comparison to other Usability-questionnaires, as the statements can be assessed even after only a short insight into the app.

In addition, the SUS can be used on small sample sizes with reliable results and is a valid measurement tool, so can effectively differentiate between usable and unusable systems.[4]

[3] http://www.tbistafftraining.info/smartphones/documents/b5_during_the_trial_usability_scale_v1_09aug11.pdf (Retrieved on 15.01.2022)

[4] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (Retrieved on 15.01.2022)

The statements and the corresponding scale of the SUS 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