Quantitative criteria and experimental procedure and setup

Quantitative criteria

For assessing the prototype subjectively and objectively, quantitative criteria were defined. These were derived from the objective and subjective measures of the individual hypotheses.

The following table shows the quantitative criteria for the individual hypotheses. The quantitative criterion always describes the threshold value above which the usability of the system can be judged as sufficient.

The criterion values for hypotheses 2.1, 2.2, 2.3, and 3.4 are taken from a study by Jeff Sauro, who found that the average task completion rate is 78%.[1] For the Craftive application, a rate of 78% and above is therefore considered satisfactory.

The criterion values for hypotheses 5.1 and 5.2 are also based on the study by Jeff Sauro. He stated, that the average number of errors per task is 0.7, with 2 out of every 3 users making an error.[1] This illustrates that it is almost impossible not to make mistakes. Thus, an error rate of no more than 70% is also an indicator of satisfactory usability for our use case.

Most of the remaining criterion values were determined by means of a test run. A person who had already acted as a test person in the qualitative analyses of the previous iterations and is therefore not a professional of the app, but who nevertheless did not see the app for the first time, carried out the tasks specified for the quantitative analysis. The values determined with the help of this test run were used to set the criterion values.

Some criterion values, such as those for hypotheses 6.1 and 6.2, were set at the subjective estimate of the app developers, always with the intention of achieving above-average performance or results.

Goals	Hypothesis	Objective measures	Quantitative Criteria
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.	Number ≥2
	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)	Number ≥ 60%
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.	Ratio (Effectiveness) ≥ 0.78

[1] https://measuringu.com/task-completion/ (Retrieved on 16.01.2022)

	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.	Ratio (Effectiveness) ≥ 0.78
	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.	Ratio (Effectiveness) ≥ 0.78
	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.	Ratio (Effectiveness) ≥ 0.78
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.	Time ≤ 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 options to fulfill this task)	The time users take to complete a specific task.	Time ≤ 40 seconds
	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.	Time ≥ 10 seconds
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.	Ratio ≥ 0.6
	4.2 On average all users can remember at least 5 system features within 30 seconds during a debriefing after the test.	The number of system features the user can remember within a specific time period during a debriefing after the test.	Number ≥ 5
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.	Ratio (Errors) ≤ 0.7
	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.	Ratio (Errors) ≤ 0.7

	5.3 The number of user errors in fulfilling all tasks will not exceed 20 on average.	Number of errors for all tasks.	Number ≤ 20
	5.4 The time spent recovering from errors will not exceed 30 seconds on average.	Time spent recovering from errors.	Time ≤ 30 seconds
	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)	Time ≤ 30 seconds
6. Subjective satisfaction	**6.1** The prototype achieves at least a rating of 6 out of 10 on average.	Rating	Value ≥ 6
	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	Number ≤ 3

The operationalisation of the SUS (System Usability Scale) was based on the statements of the usability.gov-page. These state that, based on research, a SUS score above a 68 would be considered above average and anything below 68 is below average. For the application of this system, an SUS score of ≥ 68 would be considered sufficient.[2]

Experimental Setup

- Questionnaire from LimeSurvey with the SUS (System Usability Scale)
- Zoom (shared screen)
- Link to open the App opened in the browser (mobile view activated)
- Microphone, Camera (to log facial expressions and speech)

Experimental Procedure

- Goal: test usability of the App "Craftive"
- Experiment will take place between 19.01.2022 and 30.01.2022 online via Zoom
- Each session is expected to take 30 minutes
- Needed Software: Zoom, Mobile-simulator extension for browser
- Link to the app will be shared; test user will be asked to share his/her screen during the experiment
- Opening the app by the test user will be the starting point of the experiment
- Experiment will all be conducted by one experimenter (German native speaker; to prevent linguistic distortions)
- 10 test users
- Test users will get the tasks one after the other via the chat function in Zoom → test user can read the task himself; task will additionally be read out loud by the experimenter

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

- After fulfilling all the tasks, the test user will get the link to Lime Survey with the SUS questionnaire
- No user aids beside the introduction pages of the app; experimenter will not be allowed to help
- Experiment is going to be video recorded to collect all the needed data afterwards
- Compiling all the data in an Excel File and analyzing it with SPSS

Setup the data acquisition (direct in-app logging, observation, record)

- Experiment is going to be video recorded to collect all the needed data afterwards
- Observation of the speech (regarding hypothesis 6.2)
- Time measurement for the individual tasks with the help of a timer
- Counting events (e.g. errors) by hand