Final Iteration – Quantitative Evaluation

A quantitative analysis was carried out with 10 test subjects aged 24 to 76. The exact procedure of data collection has already been described in detail in the two reports "Quantitative Criteria_Procedure_SetUp" and "Hypothesis_TestCases_Metrics" of iteration 4. But in order to better understand the subsequent quantitative data analysis, the hypotheses investigated in iteration 4 are presented again below in Table 1, including the measurement techniques used.

Table 1
Hypothesis and measurement techniques

Goals	Hypothesis	Objective measures	Quantitative Criteria	Measurement technique (variables)
1. Learnability	**1.1** All test users together use more than two ways to access a user´s profile.	The number of different commands/features to fulfill a specific task that were utilized by the user.	Number ≥ 2	Count number --> variable: number
	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%	Faster the second time? --> 3 variables: first time, second time and answer Yes (1) or No (0)
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	Manages the task? --> variable: answer Yes (1) or No (0)
	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	Manages the task? --> variable: answer Yes (1) or No (0)
	2.3 At least 78% of the test persons manage the task of agreeing on a deal	The ratio between the number of tasks completed successfully and the	Ratio (Effectiveness) ≥ 0.78	Manages the task?

	(paying for one´s help offer).	total number of tasks undertaken.		--> variable: answer Yes (1) or No (0)
	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	Manages the task? --> variable: answer Yes (1) or No (0)
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	Time to complete the task --> variable: time
	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	Time to complete the task --> variable: time
	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	Time spent --> variable: time
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	Remembers all the steps? --> variable: answer Yes (1) or No (0)

	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	Remembers at least 5 features? --> variable: answer Yes (1) or No (0)
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	Errors made during task --> 2 variables: number of errors and answer (error made): Yes (1) or No (0)
	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	Errors made during task --> two variables: number of errors and answer (error made): Yes (1) or No (0)
	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	Errors made in total --> variable: number of errors
	5.4 The time spent recovering from errors will not exceed 30 seconds on average.	Time spent recovering from errors.	Time ≤ 30 seconds	Time measurement --> variable: recovery time
	5.5 The amount of "dead" time when the user is not interacting with the system (thinking—	Amount of "dead" time (thinking-time delays)	Time ≤ 30 seconds	Time measurement --> variable: dead time

	time delays) for all tasks does not exceed 30 seconds on average.			
6. Subjective satisfaction	**6.1** The prototype achieves at least a rating of 6 out of 10 on average.	Rating	Value ≥ 6	Variable: Rating value
	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	Counting number of clear frustrations --> variable: number
7. Age	**7.1** The SUS Score is lower the older the test subjects are.	Korrelation SUS Score and Age	Korr significant and >= 0.3	Corr(Age, SUS Score) --> variable: correlation coefficient

To better understand the variable codes used in the data analysis, an overview of the variables and their coding is presented below in Table 2.

Table 2
Variable codes and descriptions

Variable Code	Variable Description
Test subject number	Number of the test subjects from 1 to 10
Age	Age of the test subjects in years
Hyp1.1	Number of the different possibilities to access a user´s profile (1 or 2)
Hyp1.2_1	Time needed for the completion of the first request in seconds
Hyp1.2_2	Time needed for the completion of the second request in seconds
Hyp1.2_3	Faster the second time? Yes(1) or No(0)
Hyp2.1	Manages the task of checking the rating of a user´s profile? Yes(1) or No(0)
Hyp2.2	Manages the task of deleting a request? Yes(1) or No(0)
Hyp2.3	Manages the task of agreeing on a deal? Yes(1) or No(0)

Hyp2.4	Manages the task of viewing the amount of their own liked projects? Yes(1) or No(0)
Hyp3.1	The time needed for creating a new request in seconds
Hyp3.2	The time needed for geeting to and liking a specific project in seconds
Hyp3.3	The time the user reads the introduction pages in seconds
Hyp4.1	Remembers all steps to fulfill the task of texting a possible helper? Yes(1) or No(0)
Hyp4.2	Remembers 5 fitures of the system within 30 seconds? Yes(1) or No(0)
Hyp5.1_1	Number of errors made while commenting a specific project
Hyp5.1_2	Errors made while commenting a specific project? Yes(1) or No(0)
Hyp5.2_1	Number of errors made while texting an intersted helper
Hyp5.2_2	Errors made while texting an interested helper? Yes(1) or No(0)
Hyp5.3	Number of errors made while fulfilling all tasks
Hyp5.4	Time spent recovering from errors in seconds
Hyp5.5	“Dead” time in seconds
Hyp6.1	Rating out of 1 to 10
Hyp6.2	Number of times the user expresses clear frustration
SUS1-SUS10	Answers of the SUS items 1-10 on a scale from 1 to 5
SUS_Score	Calculated SUS Score according to Brooke (1996)

The quantitative analysis conducted using Excel and SPSS will be presented in the following.

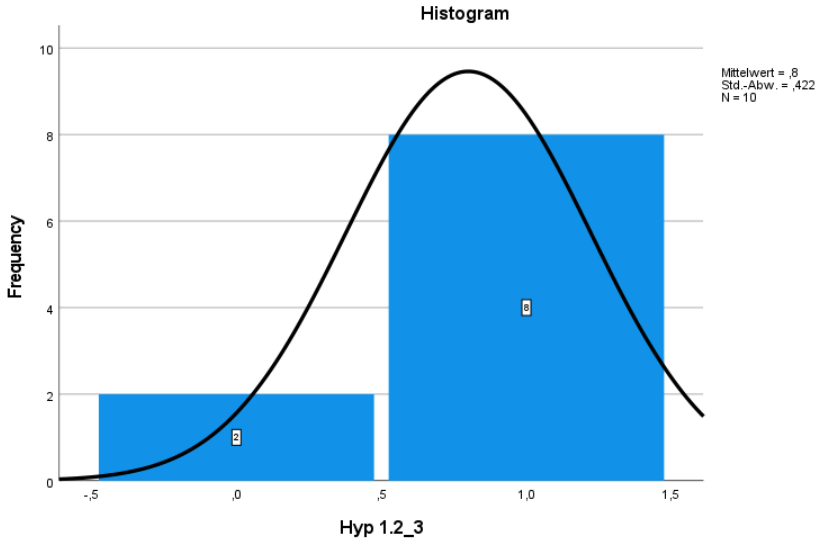
First, the correct scale levels were assigned to the data in SPSS. The variables measured by a Likert scale (Hyp6.1), as well as the SUS_Score, are considered interval scaled, i.e. quasi-metric, even though Likert scales are normally ordinal scales. However, this procedure is frequently used in statistical research and can therefore also be applied to the present case (Völkl & Korb, 2018).

To test hypothesis 1.1, the Excel spreadsheet was used to check whether both ways to access a profile were used. Since this is the case, hypothesis 1.1 can be accepted.

To test hypothesis 1.2, a frequency analysis of the variable Hyp1.2_3 was used to check the proportion of users who were faster in creating the second request, i.e. the variable is equal to 1. This was the case for 80% of the users. Since 80% is higher than the 60% required in the quantitative criteria of the hypothesis, hypothesis 1.2 can be accepted (Figure 1).

Figure 1

Histogram of the variable Hyp1.2_3



To test hypotheses 2.1, 2.2, 2.3 and 2.4, the frequencies of the variables Hyp2.1, Hyp2.2, Hyp2.3 and Hyp2.4 were tested respectively. More precisely, it was checked in how many cases the users successfully completed the task on average, i.e. the variable is equal to the value 1. For hypothesis 2.1 this corresponded to 70% of the cases, in hypothesis 2.2 60% of the cases, in hypothesis 2.3 also 70% of the cases and for hypothesis 2.4 it can be stated that 80% of the users successfully completed the task (Table 3).

Since the quantitative criteria set corresponds to a ratio of 78% so that one can speak of effective task execution, hypotheses 2.1, 2.2 and 2.3 must be rejected and only hypothesis 2.4 can be accepted.

Table 3

Frequencies of the variables Hyp2.1, Hyp2.2, Hyp2.3 and Hyp2.4

Variable		Hyp2.1	Hyp2.2	Hyp2.3	Hyp2.4
0	Frequency	3	4	7	2
	Probability	30,0	40,0	70,0	20,0
1	Frequency	7	6	3	8
	Probability	70,0	60,0	30,0	80,0
Total	Frequency	10	10	10	10
	Probability	100,0	100,0	100,0	100,0

Hypotheses 3.1, 3.2 and 3.3 were tested by looking at the mean values of the respective associated variables 3.1_1, 3.2_1, 3.3_1. This means that the average time that the users needed to complete the tasks was examined in each case. The investigation revealed an average value of 75.8 seconds for the variable Hyp3.1_1, a value of 24.4 seconds for the variable Hyp3.2_1 and an average value of 30 seconds for the variable Hyp3.3_1.

For hypothesis 3.1, the assumed quantitative criteria value is 100 seconds. Since 75.8 is smaller than 100, the hypothesis formulated in 3.1 does not have to be rejected. Since 24.40 seconds is also below the criterion value of 40 seconds formulated for hypothesis 3.2, this hypothesis does not have to be rejected either. The same applies to hypothesis 3.3 The quantitative criteria value here is 10 seconds, the value determined by means of quantitative analysis is 30 seconds, which is higher, as intended by the hypothesis.

The hypotheses regarding memorability, 4.1 and 4.2 were tested by looking at the frequency values of the variables Hyp4.1 and Hyp4.2 here as well. For hypothesis 4.1, it was determined that 60% of the users could remember the steps, so the variable value corresponded to a 1 (Table 4). The target value of 60% on average across all users could thus just be achieved, so that hypothesis 4.1 does not have to be rejected.

In the quantitative analysis of hypothesis 4.2, it was checked in what percentage of cases the variable Hyp4.2 had a value of 1, as this means that the user could list more than 5 features in the given time (Table 4). This was true in 100% of the cases, which is why hypothesis 4.2 can be accepted.

Table 4

Frequencies of the variables Hyp4.1 and Hyp4.2

Variable		Hyp4.1	Hyp4.2
0	Frequency	4	0
	Probability	40,0	0,0
1	Frequency	6	10
	Probability	60,0	100,0
Total	Frequency	10	10
	Probability	100,0	100,0

Hypothesis 5.1 and 5.2 tested what percentage of users make a mistake while performing a particular task. For this purpose, the descriptive statistics, i.e. the frequency tables of the variables Hyp5.1_2 and Hyp5.2_2 were examined again. It was found that 10% of the users make at least one mistake while commenting and 0% of the users make at least one mistake while texting. Both values are below the 70% error rate assumed in the hypothesis, which is why the two hypotheses must be rejected. However, this can be seen as positive, as it means that users make fewer mistakes than assumed.

For hypotheses 5.3, 5.4 and 5.5, the average values of the variables, here Hyp5.3_1, Hyp5.4_1 and Hyp5.5_1, were again considered. The average number of all errors made while using the app was 8, which is below the criterion value set in the hypothesis. The average value for the variable Hyp5.4_1 was 18.3, which is also below the quantitative criteria value of 30. Only the average value for the variable Hyp5.5_1 was 35.7 seconds, which is above the assumed value of 30 seconds. Thus, hypothesis 5.5 must be rejected, while the other two hypotheses mentioned can be accepted.

Hypothesis 6.1 was tested by determining the average rating value (variable Hyp6.1). This was found to be 6.1, which is higher than the target value of at least 6, and therefore Hypothesis 6.1 can be accepted as well.

In order to be able to test hypothesis 6.2, it was checked whether the average number of expressed clear frustrations was above 3. The average value of variable 6.2_1, i.e. the number of expressed clear frustrations, was 0.8 on average and thus significantly below the value of 3. Hypothesis 6.2 can therefore also be accepted.

Subsequently, the average SUS score (variable SUS_Score) was determined. This was 70.75. Since a mean score of 68 was aimed for, the score could be achieved.

In order to test hypothesis 7, a correlation between the SUS score and age was performed. First, however, the two variables Age and SUS_Score were checked for normal distribution. To do this, the two frequency distributions were first looked at graphically (Figure 2 and 3).

Figure 2

Histogram of the variable Age

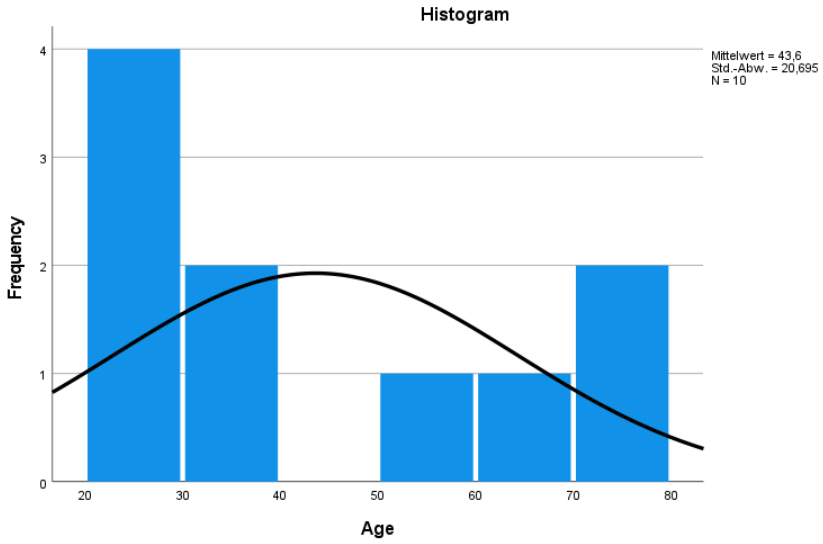
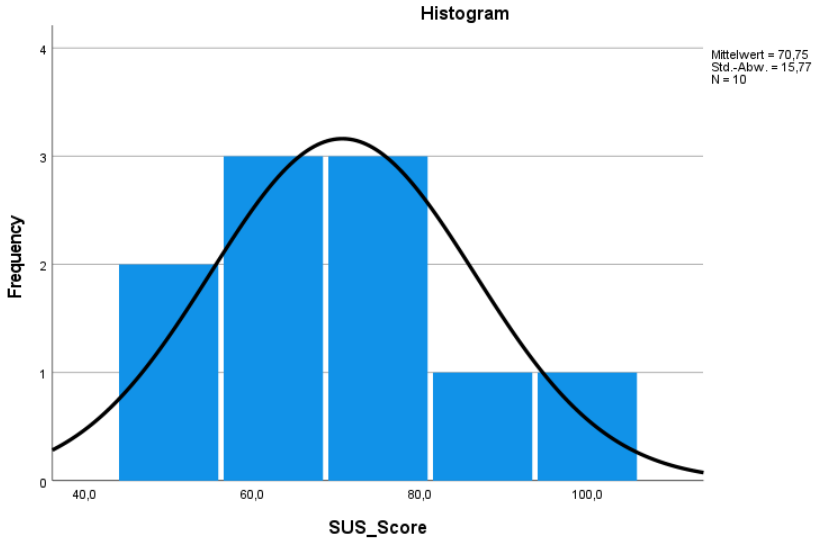


Figure 3

Histogram of the variable SUS_Score



Since these did not allow any clear conclusions, the Shapiro-Wilk test was also carried out (Table 5). However, this was not significant for either variable, which is why normal distribution can be assumed and a correlation according to Pearson could be carried out.

Table 5

Testing the normal distribution of the residuals using the Shapiro-Wilk test

Variable	Shapiro-Wilk Teststatistik W	P
Age	.845	.050
SUS_Score	.964	.832

However, the correlation of the two variables SUS_Score and Age did not result in a significant correlation (r=-.597, p=.068, n=10; Table 6), which is why hypothesis 7.1 must be rejected.

Table 6

Correlation between the variables Age and SUS_Score

Variables		SUS_Score
Age	*r*	-.597
	p	0.068
	n	10

In the following Table 7 an overview which hypotheses could be accepted and which had to be rejected can be seen.

Table 7

Overview over approved or rejected hypothesis

Hypothesis	Approved or rejected
1.1	Accepted
1.2	Accepted
2.1	Rejected
2.2	Rejected
2.3	Rejected
2.4	Accepted
3.1	Accepted
3.2	Accepted
3.3	Accepted
4.1	Accepted

4.2	Accepted
5.1	Rejected
5.2	Rejected
5.3	Accepted
5.4	Accpeted
5.5	Rejected
6.1	Accepted
6.2	Accepted
7.1	Rejected

In summary, it can be said that both hypotheses concerning learnability could be accepted, which suggests that processes within the app can be easily learned.

The execution of certain tasks described in hypotheses 3.1-3.3 could also be completed by the test subjects in a desirable time frame on average. This underlines the efficiency of the system.

The hypotheses concerning memorability could also both be accepted, which indicates that the system functions are tangible in that they can also be remembered retrospectively.

Both hypotheses 5.1 and 5.2 had to be rejected, but this is to be seen as positive, since a rejection in this case means that the test subjects made fewer errors than assumed. Two of the other hypotheses (5.3 and 5.4), which also address the error rate of the system, could be accepted, so that overall it can be stated that the error rate is relatively low despite the extensive system functions and that the users can quickly recover from errors. Only the amount of "dead" time, i.e. the time in which the user does not interact with the system because he or she is thinking, exceeded the expected time span, which must be noted negatively.

Criticism must also be levelled at the effectiveness. Three of the four hypotheses addressing effectiveness had to be rejected, as less than 78% could not successfully complete the tasks set. While 80% of the test users were able to check the number of projects liked with their own profile, they had problems deleting a request and viewing the rating details of other user profiles. Only 3 of the users were successful in making a deal with a helper, i.e. accepting their price offer and paying for it. It must also be added that some other users also discovered this function only by chance.

Nevertheless, it must be noted overall that the app still received the desired rating score of 6 on average and the number of frustrations expressed was also below the threshold of 3 on average and the desired SUS score of 68 was even exceeded.

However, there are still limitations to the quantitative analysis. One limitation, which is certainly significant, is the fact that the system was presented in English and some of the test users, especially the older ones, have only a limited command of the English language. It is possible that the test subjects would have interacted differently with the system if the app had been presented in German. In addition to this fact, the results were certainly influenced by the fact that it was sometimes difficult to measure some parameters quantitatively. For example, the amount of "dead" time or the number

of errors. It was not always clear whether the person had really made a mistake or simply wanted to try out a function out of interest.

Nevertheless, several implications for future design and functional changes of the app can be derived from the reported results. Particular attention should be paid to the effectiveness of the app. Functions such as checking the rating details of a user profile, deleting a request or especially accepting or declining an offer should be further improved. It has been shown that functions that do not contain a directly visible button but use swiping-functions instead, are often not recognised as such. In a next iteration step, the app prototype would be further improved in this respect.

References

Brooke, J.: SUS: a "quick and dirty" usability scale. In: P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Hrsg.): Usability Evaluation in Industry. London: Taylor and Francis, 1996, ISBN 978-0748404605

Völkl, K. & Korb, C. (2018). Deskriptive Statistik: Ein Einführung für Politikwissenschaftlerinnen und Politikwissenschaftler. Springer VS. https://doi.org/10.1007/978-3-658-10675-1