Usability evaluation

By Polina Anderson / Amir Dirin

Usability evaluation is a process of assessing to which degree a system is easy and pleasant to use. Usability evaluation employs various methods of a contextual nature: their effects are not deterministic and can be notably different across different projects and organisational contexts (Cockton 2014, 15.2.2).

A professional practice of evaluating usability is very diverse. The field of usability is under continuous and dynamic development; new approaches and methodologies arise with the expansion of the practice into various contexts. However, one of the possible ways to look at usability evaluation is to split the methods into 3 groups (Hom 1998):

- Usability testing: a process of testing a system with users, while carefully observing their interactions, emotional response, and behaviour.
- Usability inspection: usability assessment performed by the usability experts with no direct user involvement.
- Usability inquiry: a process of asking users for their assessment and opinions about the system. Such methods do not involve testing (first-hand observed experience with a product); rather, the methods rely on the user feedback gathered during the surveys, interviews and focus groups.

Formative and summative evaluation

Depending on the purpose and the development stage, the two types of usability evaluation are distinguished: formative and summative. According to Sauro (2015a), summative evaluation aims to determine how usable is an interface and formative evaluation inspects which elements are not usable. The terms come from educational theory, where they are serving to describe and assess student learning.

Formative or exploratory evaluation helps to form a product by measuring its usability on each development step and improving the identified issues iteratively (Rubin & Chisnell 2008, 27, 29). Formative evaluation is commonly utilized starting from the early stages of the product creation. Sometimes such evaluation is called problem discovery; it is typically conducted as a moderated user testing (Sauro 2015b).

Usability testing

Usability testing method is based on arranging an immediate user interaction with a system and analysing the results of this interaction with the help of various techniques. The method can be seen

as an irreplaceable usability practice as it gives direct input on how real users operate with the product (Nielsen 1994, 165).

According to Nielsen (2012a), usability testing could be considered as a sequence of 3 steps:

1. *Recruiting a group of representative users.*
2. *Providing the users with the representative tasks.*
3. *Observing the users performing these tasks.*

The observation typically covers user behaviour, as well as verbal and non-verbal clues expressed by the participants. The analysis, commonly driven by the predefined criteria, is heavily based on the success and failure rates of the tasks and actions performed during testing.

The essence of user testing is to let users perform the tasks on their own and to learn by observing their experience with the tested system. It is important to avoid bias in the task phrasing, the questions or moderator's behaviour in order to gather reliable uncontaminated results. Moreover, according to Nielsen (2001), it can be dramatically misleading to solely listen to what people say; a more reliable and insightful approach is to observe their interaction with the system.

As reported by Nielsen (2000), it is enough to test a particular scenario with 5 users. In most of the cases, it is more beneficial to run smaller tests and iterate them with the design revision and improvement than to conduct one big and expensive research. The best UX design practice is to start with the usability testing as soon as possible so that early design decisions are backed and supported with usability insight. Such practice allows designers to gradually address the emerging problems; it decreases the possibility to build a system based on unresolved usability issues.

System Usability Scale (SUS)

System Usability Scale is a widely applied usability measuring tool. SUS was developed by John Brooke at Digital Equipment Corporation in 1986 (Brooke 1996, 189 - 194). The crucial strength of the method is the possibility of application to any participant sample size, even a notably small one, without affecting the results (Brooke 2013, 29, 38; Sauro 2011).

The tool relies on a set of ten questions with five response options. The questionnaire is typically presented to the participants at the end of the testing session. The questions measure the overall performance of the evaluated product as well as user satisfaction. The responses are mapped on the scale from 1 to 5, where 1 is "Strongly disagree" and 5 is "Strongly agree".

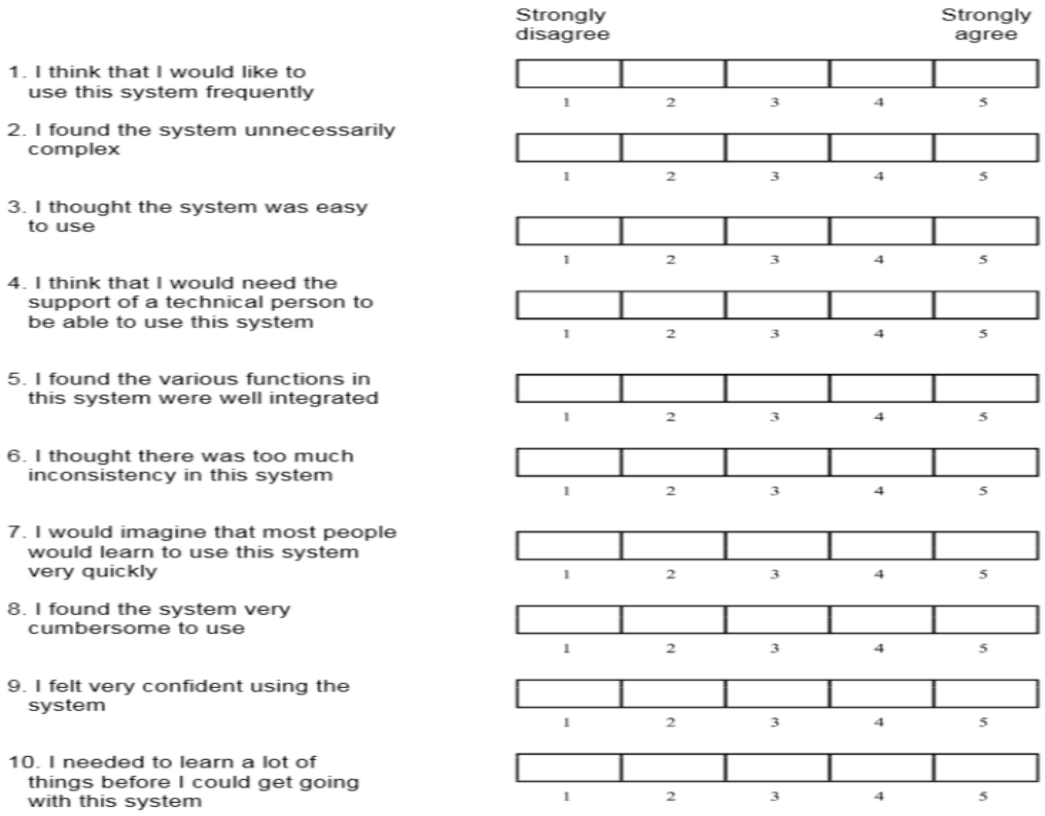

Figure 1. System Usability Scale questions (Brooke 1996, 192)

As seen from figure 1, the odd-numbered questions are formulated positively and even-numbered - negatively. Such structure is aimed to decrease bias by encouraging the users to approach the questions prudently (Brooke 1996, 191).

Upon the completion of the data collection, the results are interpreted into a single score. The score range is from 0 to 100, where the usability level increases respectively with the number. The score calculation is a rather complicated process, discussed in detail in the various literature. Typically, some of the already existing calculation tools such as spreadsheets with formulas or online apps are utilized during the analysis (Albert & Tullis 2013, 138; Sauro 2011).

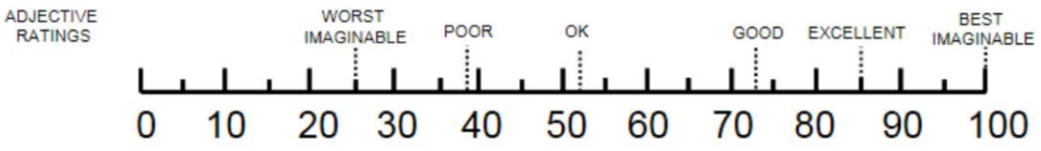

Figure 2. SUS score interpretation (Bangor, Kortum & Miller 2009, 121)

There are various ways to interpret SUS results. Bangor & al. (2009, 121) recommend using the conversion technique shown in figure 2 in order to label the usability of a product. A SUS score less than 68 is normally considered to be below average (Sauro 2011).

Although the scale is from 0 to 100, the score cannot be represented in percent as it is not changing linearly. Sauro (2011) proposed a method called “normalizing” of converting SUS score in a percentile score. The visualization of the method is presented in figure 3.

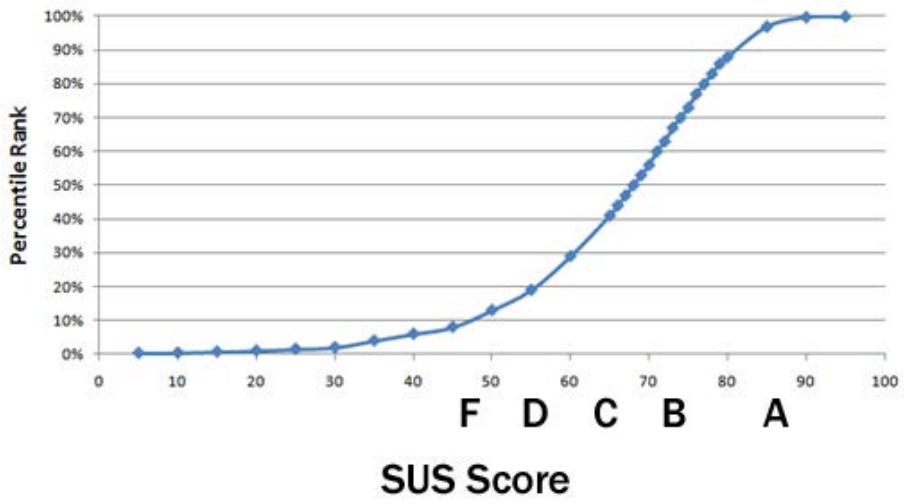

Figure 3. SUS normalizing and grading (Sauro 2011).

According to Sauro (2011), the process resembles “grading on a curve” and it is based on the distribution of all the scores. An SU score of 74 would convert to 70% on a percentile rank, whereas an SU score of 80 would be of a higher perceived usability than 80% and can be interpreted as a B - grade.

Moreover, later research brought to attention the two-factor nature of SUS. According to Lewis & Sauro (2009), it is possible to calculate not only the overall usability score but learnability and usability factors independently from each other. At the same time, the research done by Borsci, Federici, and Lauriola (2009) confirmed the two-factor structure and also showed, that the factors are correlated. The method reported by Lewis & Sauro (2009) states that learnability can be calculated by multiplying the sum SUS contribution scores from the items 4 and 10 by 12.5; usability is the sum of other items SUS scores, multiplied by 3.125. As reported by Sauro (2013a), it might be equally crucial to focus on learnability and not just on usability for some specific studies.

Usability testing framework

Nowadays, numerous ways to test usability and improve the quality of the system are available for researchers. Prior deciding upon the technique it is important to understand the goals and purposes of the planned research. Deciding which type is more suitable is a crucial decision in the entire process of usability testing (Bank & Cao 2015, 19).

It is necessary to know who the users are and what are the distinguishing factors between them, so it is possible to define user groups. The definition of the target group(s) among the user groups also supports the choice of the method. As mentioned already usability testing methods can be classified depending on the purpose (formative or summative), location (in-lab or remote) and moderator's role during the test (moderated or unmoderated).

The decisive difference between the in-lab and remote testing is that during the former the moderator and the participants are located in the same physical space. Therefore, in case of a remote testing, a moderator has to use a specialized software to communicate with participants and to record and analyse the testing session. The benefits of the remote testing over the in-lab testing are the possibility to reach larger and more geographically diverse user groups at fewer costs. It is also possible to maintain a more realistic environment for the participants and to gather more convincing usability results due to a typically larger sample of participants (Fidas, Katsanos, Papachristos, Tselios & Avouris 2007, 2-3). It is crucial that the moderator stays as objective and neutral as possible in order to not bias the participant's actions or interfere with their normal behaviour. Additionally, such roles as observers, note takers, loggers, and technicians exist to help conducting a session (Barnum 2011, 163).

Moderated and unmoderated usability testing differs mainly by the presence of a person in the role of a moderator guiding the session or the test being automatized. It is important to understand whether a potential interference in the testing process might be beneficial for the test or the aim is to get more systematic results regarding the specific area of the product.

Single Ease Question (SEQ)

Single Ease Question or SEQ is a post-task questionnaire. It assesses immediate user impressions in terms of how challenging the task was (Barnum 2011, 176, Sauro 2015c, 122). The assessment is done on a seven-point scale from "Very difficult" to "Very easy"

In-lab moderated usability testing

Traditional in-lab usability testing is one of the most frequently used evaluation method, as it has various benefits. Typically, only one participant goes through the set of tasks on a usability testing session (Albert & Tullis 2013, 53). The method requires a relatively small number of users (4 - 10) and involves personal interaction between a test moderator and a participant.

It is crucial that the moderator stays as objective and neutral as possible in order to not bias the participant's actions or interfere with their normal behaviour. Additionally, such roles as observers, note takers, loggers, and technicians exist to help conducting a session (Barnum 2011, 163).

The decisive benefit of the in-lab usability testing over the various remote options is the significantly lower “degree of separation” between the moderator and the participants. The common shared space allows to create more personal and, therefore, successful communication channels and to enhance mutual understanding. Testing in a shared location aids the observation of the facial and emotional cues of the participants. It allows to facilitate synchronous communication and to establish concepts between the moderator and the testers (Fidas et al 2007, 3).

Think Aloud protocol: concurrent and retrospective

Think aloud protocol is one of the most efficient usability testing methods, widely applied in various in-lab and remote studies (Nielsen 2012b). The essence of the method is to invite users to speak "out loud" and to vocalize their thoughts and feelings while going through a set of tasks (Rubin & Chisnell 2008, 204; Nielsen 2012b).

Two types of ‘thinking aloud’ exist in the methodology: concurrent (CTA) and retrospective (RTA). The most common approach is to employ concurrent Think aloud method - to ask users to comment on the ongoing interaction immediately and continuously (Rubin & Chisnell 2008, 204). As opposed to CTA, retrospective thinking aloud happens at the end of the test: user summarises the feedback and the impressions from the interaction with a system. The recall of the impressions is typically facilitated by observing audio/video recordings of the session.
The drawbacks of CTA are that such way of functioning as continuously commenting on the experience out-loud is not natural for humans and can cause frictions in the user performance. Some users may take time to get used to this mode, some fail to adapt at all (Nielsen 2012b). In contrary, a research done by McDonald & Petrie (2013, 2943) shows that CTA has no significant effects on performance. However, users still can find themselves frustrated and perceive the tasks as more complex than they are while performing a simultaneous thinking aloud.

Observation techniques in usability testing

Observations techniques allow gathering data by watching the participants of usability testing during the session. Observation of user interaction with a system can provide researchers with various types of feedback: verbal (language) and facial expressions, gestures, and sighs (Barnum 2011, 138).

It is advised to invite various stakeholders such as designers, managers, and developers to observe the test session, as it increases the level of acceptance of the findings (Nielsen 2010). Typically, the testing session is video/audio recorded to enable recurring access and further analysis of the gathered data.

Quantification of the usability testing results

A new approach of converting qualitative usability testing insights into quantitative data was employed in this study. The method was proposed by Latin (2017) as a way to test overall user experience and to pinpoint potential usability issues. The method provides a way to translate qualitative UX research findings into measurable and calculable data. According to Latin (2017), such approach helps to bring the UX concerns at the centre of data-driven business models.

The core idea of the method is to split the testing into small steps called scenario tasks. Upon the completion of the task (either successful or not), the participant should rate the perceived difficulty of the task on a scale from one (very easy) to seven (very difficult). This approach employs the Single Ease Question SEQ usability method. The important aspect is that all of the scores measuring difficulty are inverted SEQ scores. For example, if a task difficulty was rated as “not difficult at all” (SEQ score SSEQ = 2), the number written in the table equals 6. Therefore, the formula to calculate the score was slightly modified:

$$S_{total} = S_{success} + S_{diffivulty} = S_{success} + \left(7 + 1 - S_{SEQ}\right) = S_{success} + (8 - S_{SEQ})$$

The scores are calculated for each task. Next, they are written in a table for each session. Numerous types of analysis can be performed with such data: from calculating an average score for a task to comparing various user group performance within the testing session.

An example of Quantified results report

The video recordings, SUS data, and open-ended questions, as well as handwritten notes from the session, were employed during the analysis step. The recordings were replayed in order to extract the information needed for the measuring technique (Latin 2017), as well as the general findings and their rating among the participants. It was important to analyse how often users ran into the same usability issues and how the issues were resolved. The scores were calculated based on the success and difficulty of the tasks and were later composed into a table for each session. The tables can be found in appendix 1 (scenario 1 - 6) respectively.

Each task is marked with a code; the relevant task question can be found in appendix 2 (questions list 1). In the course of the analysis, success and difficulty rate for each task influenced the score directly. The rating scale of the variables is presented in tables 1-2.

Table 1. Success and failure score decoding

Value	Condition	Case
1	Failure	Complete failure
2	Indirect success	Failure at first, but completion in the next attempts
3	Direct success	Immediate task completion

Table 2. Measure of the difficulty of the task (SEQ score)

Code	Name of user group
0	The task was not given to a participant
1	Very easy
2	Easy
3	Quite easy
4	Medium
5	A bit difficult
6	Difficult
7	Very difficult

The data about the success and difficulty of the tasks was recorded in a table to assist usability test analysis. As explained in chapter 3.3, the formula used to quantify and calculate the scores was:

$$Score = Success\ + (8 - difficulty).$$

In table 3 the average score between all the sessions is reflected. The red numbers signalize the low average of the task (below seven). The orange numbers represent the tasks which had a low score in at least one of the sessions.

Table 3. Average score of tasks (sessions 1-6)

AVERAGE OF USER TESTING SESSIONS

Average score of tasks by the success rate and difficulty

Task	Success	Difficulty	Score
T1	3.00	1.17	9.83
T2	3.00	2.50	8.50
T3	0.83	3.17	**5.67**
T4	3.00	2.60	8.40
T5	3.00	3.50	7.50
T6	2.60	2.80	7.80
T7	0.67	5.17	**3.50**
T8	3.20	3.60	9.20
T9	2.67	1.00	9.67
T10	3.00	1.83	9.17
T11	3.00	1.20	9.80
T12	3.00	1.17	9.83
T13	3.00	1.50	9.50
T14	2.50	2.50	8.00
T15	2.67	1.00	9.67
T16	3.00	1.50	9.50
T17	2.50	2.83	7.67
T18	3.00	1.60	9.40
T19	0.20	4.40	**2.20**
T20	0.67	4.33	**3.00**
Average			7.89
Low average			3.59

The lowest score of 2.20 out of 10 was observed for the task T19 “*When is the next thesis evaluation seminar?*”. The evaluation seminar schedule for the BIT students was located on the program specific guidelines page. To a big extent, such low score is connected to the confusion around the naming and the precise meaning of the term. For the majority of the users, it was unclear whether an evaluation seminar is the same as thesis seminar and what the possible differences were. Users were not sure to what part of the process the term is related to, therefore, some of the users tried the global search and others were browsing and skimming through pages in search of a keyword.

The task T20 “To whom you should send your abstract for a language check?” scored only 3 out of 10. First of all, users were not sure what an abstract is and what kind of language check is meant - automatic or manual, by the writer or by a third-party. Mostly, students were trying to use global search, which was not very helpful. The success was primarily among the users who were patiently browsing through the step-by-step guidelines ("ABC step guide" page or a program specific guideline for BIT).

The task T7 “What are the different forms of a thesis?” got a score of 3.5 out of 10 possible points. The thesis forms are research, product-based, portfolio and diary. The reason behind the increased difficulty or indirect success of the task was that the information about the thesis types was distributed on a few pages. Users were guided to the general thesis page, where the information about only 2 out of the 4 types can be found. Some information is “hidden” in the

documents and or it is spread over pages. As a result, users either were giving an incomplete answer or were stating, that the task is difficult because they are not sure if the information they have found was comprehensive.

The task T3 “What documents do you need to sign before the beginning of the [thesis] work?” was marked with a score of 5.67. Again, the reason behind this is the dispersity of information. There is no list with all the agreements or documents and users typically spot only a commissioning agreement or a commissioning agreement with the agreement of confidential appendices, overlooking the research permit (and other possible agreements the author is not aware of). The different list of the documents could be observed from the "ABC Quick Guide" page and the program-specific guidelines. This fact is giving an impression of incompleteness or irrelevance of the found information. Users often mentioned that in such situation they would ask their thesis advisors for help.

References:

Cockton, G. 2014. Usability Evaluation. The Encyclopedia of Human-Computer Interaction, 2nd Ed. 15. URL: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/usability-evaluation. Accessed: 3 January 2018.

Hom, J. 1998. The Usability Methods Toolbox Handbook. http://usability.jameshom.com/. Accessed: 15 February 2018.

Sauro, J. 2010. If you could only ask one question, use this one. MeasuringU. Measuring Usability LLC. URL: http://www.measuringu.com/blog/single-question.php. Accessed: 16 February 2018.

Rubin J. & Chisnell, D. 2008. Handbook of usability testing: how to plan, design, and conduct effective tests. 2nd ed. Wiley Publishing, Inc. Indianapolis. URL: http://ccftp.scu.edu.cn:8090/Download/efa2417b-08ba-438a-b814-92db3dde0eb6.pdf. Accessed: 15 February 2018.

Sauro, J. 2010. If you could only ask one question, use this one. MeasuringU. Measuring Usability LLC. URL: http://www.measuringu.com/blog/single-question.php. Accessed: 16 February 2018.

Nielsen, J. 2012a. Usability 101: Introduction to Usability. Nielsen Norman Group. URL: https://www.nngroup.com/articles/usability-101-introduction-to-usability/. Accessed: 10 March 2018.

Nielsen, J. 1994. Usability Engineering. Academic Press Inc. Cambridge, Massachusetts.

Nielsen, J. 2001. First Rule of Usability? Don't Listen to Users. URL: https://www.nngroup.com/articles/first-rule-of-usability-dont-listen-to-users/. Accessed: 10 March 2018.

Nielsen, J. 2000. Why You Only Need to Test with 5 Users. Nielsen Norman Group. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Accessed: 10 March 2018.

Brooke, J. 2013. SUS: a retrospective. Journal of Usability studies, 8, 2, pp. 29-40. URL: http://uxpajournal.org/wp-content/uploads/pdf/JUS_Brooke_February_2013.pdf. Accessed: 15 February 2018.

Albert, B. & Tullis, T. 2013. Measuring the user experience: collecting, analyzing and presenting usability metrics. 2d ed. Morgan Kaufmann Publishers Inc. San Francisco.

Bangor, A., Kortum, P. & Miller, J. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. Journal of Usability studies, 4, 3, pp. 114-123. URL:

http://uxpajournal.org/wp-content/uploads/pdf/JUS_Bangor_May2009.pdf. Accessed: 24 March 2018.

Sauro, J. 2011. "Measuring usability with the System Usability Scale (SUS)". MeasuringU. Measuring Usability LLC. URL: http://www.measuringu.com/sus. Accessed: 25 March 2018.

Lewis, J. & Sauro, J. 2009. The Factor Structure of the System Usability Scale. Human Centered Design. Springer Berlin Heidelberg, 2009, pp. 94-103. URL: https://pdfs.semanticscholar.org/9cc1/bfd7d9cc4612f4d4dc15fe04edcacd95dbf1.pdf. Accessed: 24 March 2018.

Borsci, S., Federici, S., & Lauriola, M. 2009. On the dimensionality of the System Usability Scale: a test of alternative measurement models. Cognitive processing. 10 (3): 193–197. DOI:10.1007/s10339-009-0268-9.

Sauro, J. 2013a.10 Things to Know About the System Usability Scale (SUS). MeasuringU. Measuring Usability LLC. URL: http://www.measuringu.com/blog/10-things-SUS.php. Accessed: 3 February 2018.

Bank, C., Cao, J. 2015. The Guide to Usability Testing. UXPin. URL: https://www.uxpin.com/studio/ebooks/guide-to-usability-testing/. Accessed: 15 February 2018.

Fidas C., Katsanos, C., Papachristos, E., Tselios, N. & Avouris, N. 2007. Remote usability evaluation methods and tools: a survey. In Papatheodorou, T., Christodoulakis, D. & Karanikolas, N. Proceedings of PCI 2007, pp. 151-163. Patras. URL: http://hci.ece.upatras.gr/images/Papers/c126_remot_usability_evaluation_survey.pdf. Accessed: 18 February 2018.

Barnum, M. 2011. Usability testing essentials: ready, set...test! Elsevier Inc. Burlington.

Rubin J. & Chisnell, D. 2008. Handbook of usability testing: how to plan, design, and conduct effective tests. 2nd ed. Wiley Publishing, Inc. Indianapolis. URL: http://ccftp.scu.edu.cn:8090/Download/efa2417b-08ba-438a-b814-92db3dde0eb6.pdf. Accessed: 15 February 2018.

McDonald, S. & Petrie, H. 2013.The effect of global instructions on think-aloud testing. In CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2941-2944. Paris.

Latin, M. 2017. Measuring and Quantifying User Experience. URL: https://uxdesign.cc/measuring-and-quantifying-user-experience-8f555f07363d. Accessed: 20 March 2018.

Appendix 1.

Appendix 6. Scores for each task for the usability testing sessions

Session 1. Task scores

SESSION 1

Average score of tasks by the success rate and difficulty

Task	Success	Difficulty	Score
T1	3.00	1.00	10.00
T2	3.00	1.00	10.00
T3	1.00	2.00	7.00
T4			0.00
T5	3.00	3.00	8.00
T6	3.00	2.00	9.00
T7	0.00	5.00	**3.00**
T8	3.00	5.00	**6.00**
T9	3.00	1.00	10.00
T10	3.00	3.00	8.00
T11	3.00	1.00	10.00
T12	3.00	1.00	10.00
T13	3.00	1.00	10.00
T14	3.00	3.00	8.00
T15	3.00	1.00	10.00
T16	3.00	1.00	10.00
T17	3.00	2.00	9.00
T18	3.00	1.00	10.00
T19	0.00	5.00	**3.00**
T20	1.00	6.00	**3.00**
Average			8.11
Low average			3.75

Session 2. Task scores

SESSION 2

Average score of tasks by the success rate and difficulty

Task	Success	Difficulty	Score
T1	3.00	1.00	10.00
T2	3.00	1.00	10.00
T3	1.00	2.00	7.00
T4	3.00	1.00	10.00
T5	3.00	4.00	7.00
T6			0.00
T7	0.00	7.00	**1.00**
T8	1.00	1.00	8.00
T9	3.00	1.00	10.00
T10	3.00	1.00	10.00
T11	3.00	1.00	10.00
T12	3.00	1.00	10.00
T13	3.00	3.00	8.00
T14	3.00	1.00	10.00
T15	3.00	1.00	10.00
T16	3.00	2.00	9.00
T17	3.00	1.00	10.00
T18	3.00	1.00	10.00
T19	0.00	7.00	**1.00**
T20	3.00	1.00	10.00
Average			8.47
Low average			1.00

Appendix 2. Usability testing questions

Questions list 1. Usability testing scenario question

Task code	Question
T1	You need information about thesis writing process in Haaga-Helia. Please, try to find guidelines for it.
T2	Please find out the steps of the thesis process you need to follow.
T3	What documents you need to sign before the beginning of the work?
T4	Can you find a template for your thesis?
T5	Find a thesis coordinator's name of your program
T6	Find a thesis coordinator's name of MUBBA's program
T7	What are the different forms of thesis?
T8	What would be your steps during the thesis writing?
T9	To whom you can write, if you need help?
T10	Investigate the steps to take when your thesis is written already? What do you need to do to finalise the process?
T11	Check, how will your thesis be graded? (Scale, areas etc.)?
T12	How to submit your thesis for Urkund check?
T13	Where can you get a voice recorder or other equipment in order to conduct an interview for your thesis?
T14	Whom you should contact, if you want to get information about "Mission Goes Global" thesis group?
T15	What are the options of the databases for publishing your thesis, when it will be done? What's the difference between them?
T16	In what language should you write your maturity examination? What are the options?
T17	Is it possible to retake a maturity test?
T18	Find a thesis topic, proposed by a company for Haaga-Helia's students.
T19	When is the next thesis evaluation seminar?
T20	To whom you should send your abstract for a language check?

Questions list 2. The two open-ended question in the post-test

1. At the beginning of the test, the following question about MyNet thesis pages was asked: "What tasks it's possible to perform with these pages"? As you now have gone through almost all the functions of the pages, please assess the level of the exposed essential information available from the first sight. Which important content needs more attention and should be exposed more properly during the first browsing of the pages?
2. Please, give us the final feedback about the product you have just tested. How do you feel about it, what emotions you had during the testing and what are your impressions?

Questions list 3. The two open-ended question in the post-test

1. I think that I would like to use this portal frequently.
2. I found the portal unnecessarily complex.
3. I thought the portal was easy to use.
4. I think that I would need the support of a technical person to be able to use this portal.
5. I found the various functions in this portal were well integrated.
6. I thought there was too much inconsistency in this portal.
7. I would imagine that most people would learn to use this portal very quickly.
8. I found the portal very awkward to use.
9. I felt very confident using the portal.

10. I needed to learn a lot of things before I could get going with this portal.