



Utrecht University

FACULTY OF SCIENCE
HUMAN COMPUTER INTERACTION
MASTER'S THESIS

AUGUST 2023

**From Fragmentation to Uniformity:
Towards a Standardized Decision Process
for UX Evaluation Methods in a Large
End-To-End Agency**

Supervisors:
dr. Christof van Nimwegen
Imke de Jong, MSc

Student:
Annemik Stolk, BSc
6173055

Abstract. This thesis demonstrates a proof of concept of a standardized decision process for user experience evaluation methods (UXEMs). A large end-to-end agency serves as a case study for this thesis. The final deliverable presented in this thesis is a prototype of a mobile decision tool called DEPRO (DEcision PROcess). A multivocal literature review (MLR) was conducted. Contextual inquiry took place at the company and interviews were conducted with eight members of the user experience teams across multiple locations. The MLR yielded a list of 126 user experience evaluation methods, divided into seven categories: expert evaluations, field studies, interviews, measurements, scales & questionnaires, software tools and workshops. Important aspects of these methods were identified, as well as their general application: in academia or practice? The input from the MLR, contextual inquiry and interviews was used to create DEPRO, a mobile UXEM decision process tool. This tool was evaluated through interviews with members of the UX teams. This evaluation showed that the team members were interested in using the tool if it were a functional product and that they would use it to inform their decision.

Table of Contents

1	Introduction	1
1.1	Research questions	3
2	Methods	4
2.1	Methods overview	4
2.2	Process overview	5
2.3	Processes A & B: data collection and data analysis	9
2.4	Process C: implementation	9
2.5	Process D: testing and evaluation	9
3	Literature review protocol	11
3.1	Research questions	12
3.2	Search keywords	12
3.3	Inclusion and exclusion criteria	13
3.4	Snowballing	13
3.5	Data extraction process	14
4	Results: Literature review	15
4.1	SQ1: What types of user experience evaluation methods currently exist in both academic and corporate settings?	15
4.2	SQ2: What are the positive and negative aspects of each type of user experience evaluation method?	20
4.3	SQ3: Which of the found user experience evaluation methods are best suited for academic purposes and which are best suited for corporate purposes?	22
4.4	SQ4: What decision processes for user experience evaluation methods currently exist in academic and corporate settings?	23
5	Results: Contextual inquiry	26
5.1	Current situation	26
5.2	Current set of evaluation methods	29
5.3	Requirements	29
6	Results: Interviews	31
6.1	Way of working	31
6.2	Methods	35
7	Implementation	37
7.1	Mobile application	37
7.2	Requirements: MoSCoW	38
7.3	Content	39
7.4	Prototype	41
8	Testing and evaluation	47
8.1	Set up	47
8.2	Results	47
8.3	What now?	48
9	Discussion	49

9.1	Methods	49
9.2	Results	49
9.3	Deliverable	49
9.4	Implications	50
9.5	Future work	50
10	Conclusion	51
A	MLR UX evaluation methods	61
B	Informed consent	85
C	Interview protocol: Current situation and methods	86
	C.1 General	86
	C.2 UXEMs	86
	C.3 Current process	86
	C.4 End	86

1 Introduction

User experience (UX) is important. It is so influential that products with the best functionality may very well be outperformed by a product with better user experience [51]. Just look at the iPhone: the user experience in Apple's products weighs heavily in consumers' considerations and may be the deciding factor to choose their device. User experience cannot be ignored when designing a product, as it influences the quality of a product that customers use in their daily life. Improving a product may thus improve customers' quality of life [75]. Väänänen-Vainio-Matilla, Roto and Hassenzahl argue that user experience should be a key concern of product development, as products should be enjoyable and support fundamental human needs and values [120].

The value of user experience has become more and more recognized by the industry. Companies focus on UX, from independent entrepreneurs to large agencies. User experience experts or interaction designers are highly sought after. In the Netherlands, many companies exist that focus on user experience design and evaluation. Other companies know the value of user experience and have user experience expertise in-house, available whenever necessary. Something that happens in the industry is larger companies buying smaller companies, often with the aim of expanding their services and knowledge of certain topics.

In the industry a focus on user experience is often something that needs to be emphasized. Especially people who may not concern themselves with user experience on a daily basis may not immediately see the value user experience research can add to a product. In the field of user experience in the Netherlands, several types of players can be identified. First, the independent entrepreneurs, who own a business in user experience consultancy or design. These entrepreneurs have often worked in the field and have gained experience, after which they took matters into their own hands. Second are small companies with multiple employees with a specific focus on user experience. These two types of companies are often hired by medium sized companies who do not possess the knowledge to design or evaluate user experience in-house. Finally there are large agencies that offer user experience as a service without it being the sole focus of the company. These types of companies will often offer different but related services, for example digital marketing. They typically take on large clients for projects that are longer term.

A fusion of multiple smaller companies into a larger brand brings along many challenges. Different companies may have different ways of working and different company cultures exist. The larger an organization is, the more challenging it becomes to centralize data and knowledge. There is a substantial amount of information present in all teams, but centralized cognition is lacking. This means that every company is familiar with a certain set of methods that are not guaranteed to match those of other companies. In other words, a lack of uniformity exists within the company. The same project may have different outcomes depending on who delivers it. This may take away value for the customer, as the optimal outcome cannot be guaranteed. Additionally, an information exchange often occurs verbally as opposed to being written down [135]. This slows centralized cognition, as there is no way of retracing past conversations and thus important decisions regarding user experience evaluation methods. It may also cause

inconsistencies or misunderstandings as information is transferred and may risk being transformed every time it is discussed.

The discussed phenomenon is currently a challenge at the company that is used as a case study for this thesis. The company, a large name in the industry, is an international end-to-end agency. They offer a wide range of services to their customers to help them reach their business goals. The company has bought multiple smaller marketing and IT start-ups and companies across Europe. Based on proximity, these smaller companies are combined into several locations or hubs spread across The Netherlands, Belgium and other European countries. Four overarching branches exist within the company: technology, content and production, strategy and marketing. Every team operates within one of these branches. The user experience teams are commonly classified as content and production, but this can differ based on the expertise present within the teams. UX teams who have technical experience may fall under the technology branch. With many different companies and teams united under one large agency and many different ways of working, it is no wonder that centralized cognition is halted.

There are many different methods that can be applied to achieve the best user experience. Vermeeren et al. [127] explored 96 methods and performed an analysis based on, among other criteria, the product development phase and the studied period of experience. In their work, they differentiate between user experience and usability based on objectivity. Usability usually measures concrete aspects such as number of errors, clicks or task execution time. User experience (UX) is much more subjective, as the goal is to find out how the user feels about the system. This subjectivity is also reflected in the ISO9241-210:2019 standard definition, where it is stated that user experience can be defined as a *"user's perceptions and responses that result from the use and/or anticipated use of a system, product or service"* [61]. Väänänen-Vainio-Mattila, Roto and Hassenzahl, too, emphasize that UX in general is a subjective term [120]. UX can also be seen as an extension of the satisfaction aspect of usability [137]. Vermeeren et al. view usability as a part of UX. Following the distinction made in Vermeeren et al., two evaluation methods are identified: usability evaluation methods and UX evaluation methods [127]. As UX is a newer term, the literature often uses the term usability for something that actually represents UX.

User experience can be designed, but also evaluated. An important distinction that is made in the work of Vermeeren et al. is the one between evaluation and design methods. Design methods are intended to spark inspiration for new products or designs [45], where evaluation methods are used to help choose the best design, confirm the development is going in the right direction or to assess whether targets are met [127].

Some methods are typically applied in academia, others are more commonly seen in the industry. There are some methods that are applied more than others, but it is not always clear which method is best applied in a certain scenario. In the case study for this thesis, this is exactly the question. When decentralized cognition and a lack of uniformity exist, a decision to choose a certain method for a project cannot be properly validated or motivated to peers. Colleagues should be able to retrace the steps taken to achieve a final product to ensure an optimal outcome. A way to structure and optimize

the user experience methods decision process is required. The present thesis proposes a decision tool to aid with the structured selection of user experience evaluation methods and offers a theoretical background on the state of the art of user experience evaluation methods and decision tools through a multivocal literature review. With the addition of contextual inquiry and interviews at the aforementioned company, a proof of concept for a decision tool called DEPRO (DEsicion PROcess) is proposed.

1.1 Research questions

This thesis focuses on developing a decision aid for the process of selecting user experience evaluation methods. A prototype of an interactive mobile decision tool is envisioned as a final deliverable. The following research questions are proposed to shape this tool.

Main research question:

- RQ1: How can challenges that come with choosing user experience evaluation methods in a non-uniform way in a large end-to-end agency be solved through an interactive decision model?

To answer this main research question, several sub research questions are identified:

- SQ1: What types of user experience evaluation methods currently exist in both academic and corporate settings?
- SQ2: What are the positive and negative aspects of each type of user experience evaluation method?
- SQ3: Which of the found user experience evaluation methods are best suited for academic purposes and which are best suited for corporate purposes?
- SQ4: What decision processes for user experience evaluation methods currently exist in academic and corporate settings?
- SQ5: What are the requirements for an interactive decision support system intended to aid the decision process for user experience evaluation methods?

2 Methods

In this section, the methods used in this thesis are elaborated upon. In section 2.2 an overview of the process is provided, along with the input and output of every phase. In sections 2.3, 2.4 and 2.5, the general phases of this thesis project are explained.

2.1 Methods overview

In this thesis, there are four ways in which information is retrieved:

1. Multivocal literature review (MLR);
2. Contextual inquiry;
3. Interviews;
4. Evaluation.

Some sub research questions are answered by the literature, while others are answered through contextual inquiry and interviews. SQ1 should uncover the range of existing UX evaluation methods. SQ2 covers the positive and negative aspects of these methods. With SQ3, the purpose of the methods is investigated. SQ4 should shine light on existing decision processes concerning UX. These four sub research questions are (partially) answered through the multivocal literature review. SQ5 is meant to uncover the requirements for the final deliverable. The interviews and contextual inquiry are performed to answer (part of) all sub research questions. Finally, an evaluation of the deliverable should aid in answering the main research question.

Multivocal literature review A systematic literature review or SLR can be applied for transparent reporting of scientific literature by avoiding any biases the researcher may possess [94]. The systematic nature of this type of literature review ensures reproducibility. The SLR is applied to academic literature and thus search engines such as Scopus, Google Scholar or Web of Science may be used. For the scope of this thesis, collecting academic sources is not enough as the main focus is to investigate existing UX evaluation methods in corporate as well as academic settings. For this reason a multivocal literature review (MLR) is used. In an MLR, gray literature is reported as well.

Contextual inquiry Contextual inquiry is a method first described by Beyer and Holtzblatt [16]. It is a type of field study where in-depth observation and small-sample interviews are the main focus [109]. Contextual inquiry consists of two concepts: the context, so the environment of the user, and inquiry, the researcher watching the users perform their tasks and ask for information to understand their actions. In this case, the context is the office of location A and the inquiry is done by the author through observation and short interviews with present employees of the UX teams. The researcher plays the role of an apprentice, while the participant plays the role of a teacher. Contextual inquiry is a method that can be applied in the early discovery stages of a project.

Interviews In an interview, a researcher and a user engage in an interview about a topic of interest [99]. As the researcher, it is important to be well prepared and, in the case of semi-structured or structured interviews, to have an interview protocol prepared. For the purpose of this thesis, semi-structured interviews are conducted. Semi-structured interviews allow for elaboration when the answer to a question requires specification.

2.2 Process overview

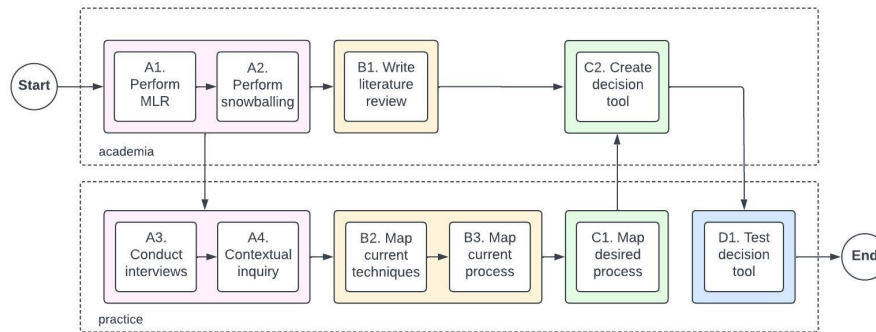


Fig. 1: Overview of the process

In Figure 1 the general steps of the process in this thesis are stated. Processes A1, A2, B1 and C1 occur separate from the company in the case study, where processes A3, A4, B2, B3, C2 and D1 require the expertise of the employees at the company in the case study in some way. In this thesis, the four most important processes that produce outcomes are the multivocal literature review, interviews, contextual inquiry and the evaluation. Four phases can be identified, represented by the colours and letters in the model below: data collection (pink), data analysis (orange), implementation and testing (green) and evaluation (blue). The figure is explained in the following sections, phase by phase.

Data collection Processes A1, A2, A3 and A4, shown in pink in figure 2, represent the data collection phase of this thesis. In A1, the multivocal literature review (MLR) is conducted, and in A2 the sources from the MLR are used to perform backward and forward snowballing. These techniques are used to gain insight in the state of the art of UX evaluation methods and decision processes.

In A3, semi-structured interviews with eight different UX team members (two from each location) are conducted. In these interviews the UX team members are asked to describe the current process from start to finish, focusing on the decision process. The techniques that they use are also a focus point in these interviews. The informed consent for these interviews can be found in Appendix B, the interview protocol can be found in

Appendix C. Interviewees are assigned a participant number to ensure anonymity and are asked prior to the interview to fill out a form stating they have read and understood the informed consent. Interviews preferably take place in real life, but depending on the availability of team members online interviews may be arranged. The interviews are annotated as the interview progresses, and with the permission of the interviewee an audio recording is made. This recording can be transcribed using Nvivo, software commonly used for qualitative analysis.

Aside from the interviews, contextual inquiry also takes place in A2. Here, company culture is observed. The data these processes yield is then used in the next phase. In table 1 the input and output of the data collection phases are shown.

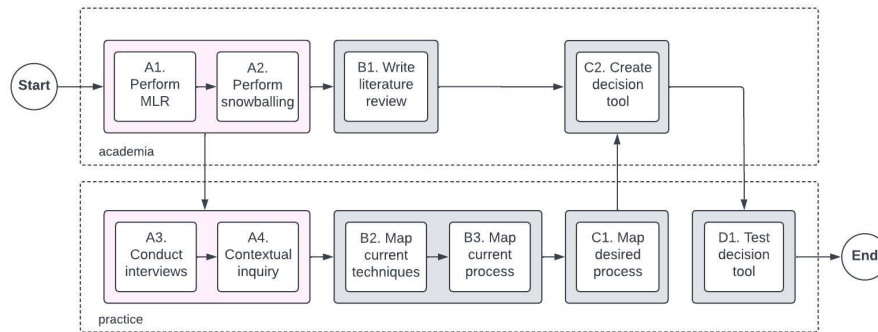


Fig. 2: Process A: data collection

Table 1: Steps of the data collection phase

Process	Input	Output
A1. Perform MLR	Keywords	MLR table
A2. Perform snowballing	MLR table	Snowballing table
A3. Conduct interviews	Interview outline and protocol	Transcript of interviews
A4. Contextual inquiry	Presence at the company	Additional context to interviews

Data analysis Processes B1, B2 and B3, shown in orange in figure 3, represent the data analysis phase of this thesis. In B1 the results from step A2 and A3 are used to write a literature review. In B2 the results from the interviews in A1 are used to map the current techniques that are present within the company. Then, the current process of deciding on a UX evaluation method is mapped in B3. In table 2 the input and output of the data analysis phases are shown.

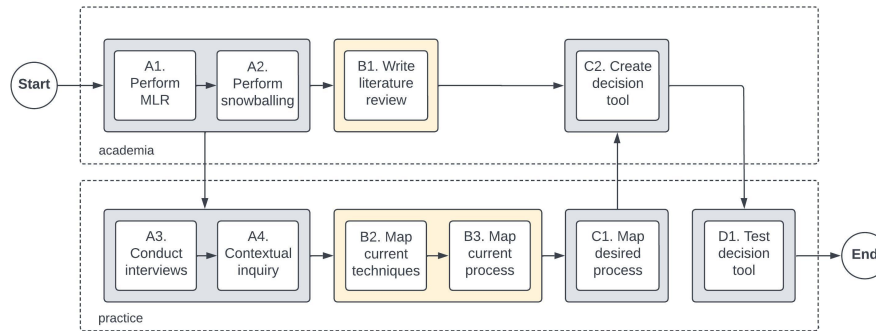


Fig. 3: Process B: data analysis

Table 2: Steps of the data analysis phase

Process	Input	Output
B1. Write literature review	MLR table; snowballing table	Literature review
B2. Map current techniques	Transcript of interviews; additional context	List of current techniques
B3. Map current process	Transcript of interviews; additional context	Model of current decision process

Implementation Processes C1 and C2, shown in green in figure 4, represent the implementation phase of this thesis where the knowledge and contexts gathered in the previous processes are implemented so a decision tool can be created. This decision tool is a prototype for a mobile decision tool to aid in user experience evaluation method selection. The prototype serves as a proof of concept. In C1 the desired process is mapped based on the current process and the techniques the company is currently using. Additional interviews with a team manager may be planned to gain some more insight in the desired situation. In C2 the first version of the decision tool is created based on the findings in the literature. This decision model is an interactive model, assisting UX team members in finding the best UX evaluation method for their specific project. The tool is somewhat comparable to concepts found in the literature [88,43]. The outcome of the decision tool is a list of UX evaluation methods, along with information about them regarding required users, time, funds and more. Sometimes one evaluation method may not unravel the whole picture, so a top list allows UX professionals to make an informed decision. A decision process model is also created to show how the interactive tool works. In table 3 the input and output of the implementation phases are shown.

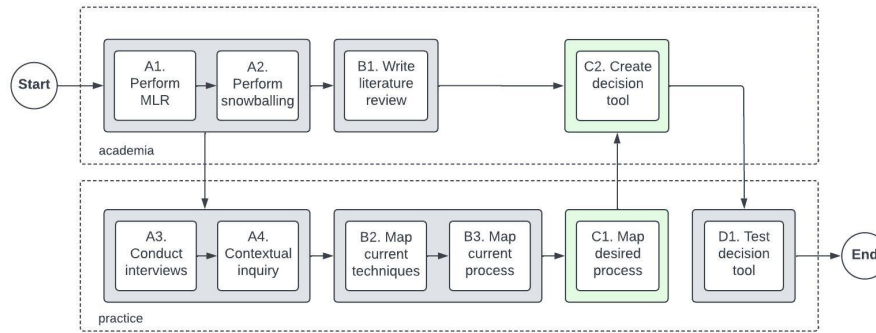


Fig. 4: Process C: implementation

Table 3: Steps of the implementation phase

Process	Input	Output
C1. Map desired process	List of current techniques; model of current decision process	Model of desired process
C2. Create decision tool	Literature review	Interactive decision tool

Testing and evaluation Process D1, shown in blue in figure 5, represents the final testing and evaluation phase of this thesis. Here the decision model is tested with members of the UX team based on a realistic case study. Based on the results of D1 the decision model may be improved and evaluated again and/or suggestions for future work may be given. In table 4 the input and output of this phase are shown.

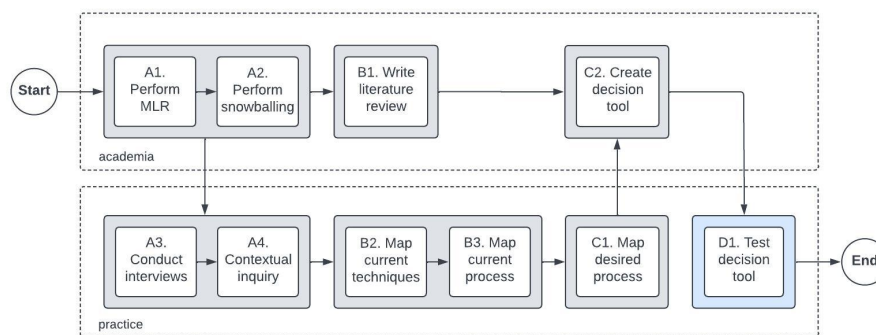


Fig. 5: Process D: testing and evaluation

Table 4: Steps of the testing and evaluation phase

Process	Input	Output
D1. Test decision model	Decision model	Improved decision model; suggestions for future research

2.3 Processes A & B: data collection and data analysis

In this thesis, contextual inquiry is used to explore the existing situation in the end-to-end agency from the case study. Where possible, existing processes are observed. The availability of new projects may influence the possibility to do so. If observation is not possible, interviews where the decision process is explained by an experienced member of a UX team may be a suitable replacement. Separate interview sessions are also held with UX team members to gain insight in the existing UX evaluation methods within the agency. All interviews are semi-structured to allow space for follow-up questions and elaboration where necessary.

The outcome of the literature review, with input gathered through an multivocal literature review and snowballing, can be compared to insights gained through contextual inquiry for SQ1 and SQ2. This way, any gaps between literature and practice can be identified. Through observation and interviews, SQ4 and SQ5 can be answered as well.

2.4 Process C: implementation

The process discussed in the sections above finally lead to the implementation of the deliverable. This deliverable is a functional prototype of a decision tool in the shape of a mobile application. The literature study provides context and information to base the decision tool on. The interviews and contextual inquiry shine light on the current situation and lead to requirements for the system. The prototype is created in Figma, as this is a tool that is familiar to the company.

2.5 Process D: testing and evaluation

In the testing and evaluation phase, the final decision tool is tested in practice in a summative way. With the help of the UX team manager of location A, a realistic customer case is created. This case may be modified from a real customer case. The case is then shown to a UX team member in a one-on-one conversation with the researcher. First, they are asked to think aloud and explain how they would normally tackle this case, focusing on the decision process of selecting a suitable UX evaluation method. Once finished, they are given the decision model and asked to apply it in the same manner, so thinking aloud. Finally, they are asked what their thoughts on the model are. This way, the outcome of both processes can be compared and based on this we can determine the success of the decision model. The time it takes to reach an outcome is measured to assess the potential added value for the general UX process. The number of steps taken to reach a decision is measured to assess the efficiency of the model. Finally, after the

evaluation, the participant is asked to fill in a small questionnaire about their experience with both approaches.

The within-subject approach is best suited for this case study as UX team members heavily rely on their own experience and often are required to make decisions by themselves. Additionally, within-subject approaches require less participants than between-subject approaches, and there is a limited number of qualified participants. Individual answers to the first half of the customer case may differ from team member to team member. The evaluations are audio recorded only if the team member has given explicit written consent.

3 Literature review protocol

In this section, the protocol for the literature review is discussed. This is a part of the *data collection* process in this thesis.

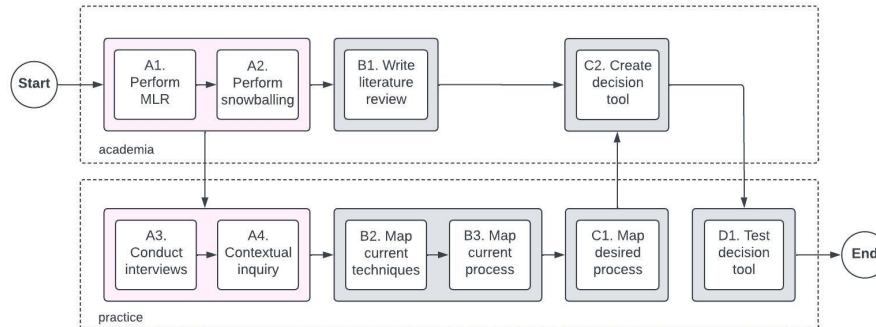


Fig. 6: Process A: data collection

A systematic literature review (SLR) focuses on academic literature. Performing an SLR can aid in avoiding researcher bias, increases reproducibility of the research and allows for transparent reporting of the literature [94]. For the purpose of this thesis, it is not just important to consider academic literature, but literature from practice as well. For this reason a multivocal literature review (MLR) is most suitable for this thesis, as it is possible to consider gray literature as well.

The boundaries of gray literature are often unclear and it is difficult to clearly define the term and all types of sources that may fall under it [9]. In 2010, Schöpfel proposed the Prague definition: *"Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body"* [111]. Cirkovic identified common characteristics of gray literature: *"difficult to identify, access and locate; often come in the form of limited editions; inaccessible in bookstores; lack of bibliography registration; absent in library collections and catalogues and in a publisher's catalogues as well; difficult to acquire in libraries; tend to be unpublished or published with delay"* [26].

As the SLR PRISMA guidelines [94] do not suffice in the case of this thesis, we need guidelines that take the broader nature of an MLR into account. In 2019, Garousi, Felderer and Mäntylä proposed guidelines for conducting multivocal literature reviews in software engineering that can be applied to other fields as well [44]. In their work, they go over existing MLR guidelines and propose a set of guidelines to ensure a

high quality of MLR processes as well as their results. In this thesis, the guidelines by Garousi, Felderer and Mäntylä are used.

For this thesis, snowballing is used as an additional search method in the MLR according to guidelines for including both backward and forward snowballing in a systematic literature review by Wohlin [133]. Any modifications to the guidelines that are required for the MLR process are documented in this section.

3.1 Research questions

The MLR aims to answer the following research questions:

- SQ1: What types of user experience evaluation methods currently exist in both academic and corporate settings?
- SQ2: What are the positive and negative aspects of each type of user experience evaluation method?
- SQ3: Which of the found user experience evaluation methods are best suited for academic purposes and which are best suited for corporate purposes?
- SQ4: What decision processes for user experience evaluation methods currently exist in academic and corporate settings?

As SQ5 is specific to the company, this sub research question cannot be answered through the MLR. Instead, this question is answered through contextual inquiry and interviews in later chapters.

3.2 Search keywords

The following key words were used while performing the MLR:

- (user experience OR UX OR usability) + (evaluation methods OR testing methods OR validation methods OR methods);
- (user experience OR UX OR usability) + (method selection OR decision OR decision process OR evaluation decision).

For the purpose of the present thesis, usability is used as a search term for the MLR and included when the meaning is relevant to UX evaluation methods. The two terms are often used in the same context. The scope of this thesis is UX evaluation methods, so this MLR is focused on UX and usability evaluation methods as opposed to design methods.

Search engines Google Scholar and Scopus (both sorted by relevance) were used to find scientific literature. Google Scholar was used to avoid bias in favour of a specific publisher [133], in this case Elsevier (Scopus). Google and DuckDuckGo were used to find gray literature. The keywords were used in all four search engines.

3.3 Inclusion and exclusion criteria

A source was accepted in the MLR when it met the following inclusion criteria:

- Source qualifies as scientific literature or gray literature.
- Source contents center around user experience evaluation methods and/or user experience decision making processes.
- Source was written or published in 2010 or more recent.
- Source is found in the first 40 results of the search engine.

A source was rejected from the MLR when it did not meet all inclusion criteria or when it met one of the following exclusion criteria:

- Source is written in any language but English.
- Source contains a significant number of spelling or grammatical errors, lowering the perceived quality of the source.
- Source is about anything other than user experience evaluation methods and/or decision making processes.

3.4 Snowballing

Only papers that passed the inclusion and exclusion criteria were eligible for the snowballing process.

Backward snowballing The first type of snowballing used in this process was backward snowballing. Here, new sources were found based on the reference list of the eligible paper. The steps of the backward snowballing were:

1. Exclude papers in reference list that do not meet the inclusion and exclusion criteria stated in section 3.3.
2. Identify duplicates (i.e. papers that have already been examined).
3. Find in-text reference to examine context of paper.
4. Find paper, study abstract, browse through paper.
5. Read paper.

At any step in the process, papers could be included or excluded.

Forward snowballing In forward snowballing new sources were found by looking at places where the eligible paper is cited. Wohlin [133] suggests using Google Scholar. For this thesis, both Google Scholar and Connected Papers were used. The steps for forward snowballing were:

1. Exclude papers based on the information provided by Google Scholar and/or Connected Papers.
2. Find paper, study abstract.
3. Find place where source was cited to examine context.
4. Study full text.

At any step in the process, papers could be included or excluded. Both the process for backward as well as forward snowballing were based on the guidelines by Wohlin [133]. Both processes were executed iteratively until no new papers were found.

3.5 Data extraction process

All sources found in the MLR process were kept in an Excel file and were then judged based on the inclusion and exclusion criteria. From each source, the following information (if available) was extracted and included in the Excel file:

- Author(s) or organization;
- Year of publication;
- Title;
- Source classification (scientific/gray literature or neither);
- Where found;
- Type of UX evaluation method described;
- Positive or negative aspects of UX evaluation method;
- Academic or corporate purpose;
- Short description of decision process;
- Link to article;
- Accepted or rejected + reason.

During the snowballing phase in section 3.4 the full paper was examined before a final conclusion on inclusion or exclusion was reached. Thus, the data extraction process could occur simultaneously.

4 Results: Literature review

In this section, the first four sub research questions are elaborated upon based on the multivocal literature review that was conducted and described in section C. Performing the MLR is part of process A: *data collection* and the literature described in this section is part of process B: *data analysis*. This is illustrated in figure 7.

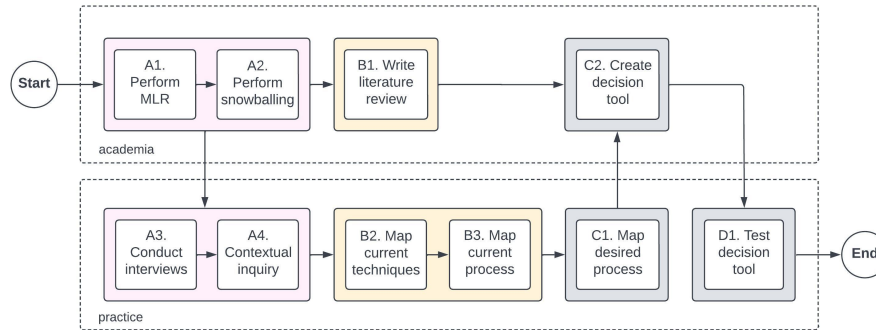


Fig. 7: Process A and B: data collection and analysis

4.1 SQ1: What types of user experience evaluation methods currently exist in both academic and corporate settings?

Paz and Pow-Sang [98] present a systematic mapping review of usability evaluation methods in the context of software development processes and investigated how many times a usability evaluation method was mentioned in the relevant papers. In total, they mention 34 usability evaluation methods for software systems in their paper, along with their definitions according to the literature.

Rico-Olarte, López and Kepplinger [106] created a conceptual framework aimed to identify differences between UX evaluation perspectives and their measurable aspects. They focused on two perspectives during the analysis: the system and the user. The creation of the framework lead them to a definition of an objective UX evaluation method, as they describe that physiological signals are the convergence point between physical state of the user and the measurement of their emotions.

In their work, Zarour and Alharbi [137] propose a theoretical UX framework with four UX dimensions: value, brand experience (BX), user needs experience (NX) and technology experience (TX). The dimensions relate to certain UX aspects. The aspect brand corresponds to BX, the aspects pragmatic and hedonic relate to NX and user experience designs, development technology, hardware and operation relate to TX. This framework could aid in classifying the found UX evaluation methods into overarching categories. Another possibly helpful distinction is made by Bernhaupt [13], who differentiates between user-oriented methods and expert-oriented methods, with a third

category called other approaches. Additionally, Kurosu, Hashizume and Ueno [75] distinguish real-time methods and memory-based methods.

Vermeeren et al. [127] use a system to represent method suitability where they rate each evaluation method on the following: study type (field studies, lab studies, online studies or questionnaires), development phase (concepts, early prototypes, functional prototypes or products on market), studied period of experience (before usage, snapshots, an episode or long-term UX), evaluator/info provider (UX experts, one user at a time, groups of users, pairs of users), data (qualitative, quantitative or both), applications (web services, PC software, mobile software, hardware designs or other) and requirements (trained researcher or special equipment). This type of suitability list may prove useful for the final decision tool.

Lachner et al. [76] identified five evaluation clusters. The first cluster is measuring sensation, where participants are presented with visuals as opposed to verbal measurements. An example of this is Emocards [32], where the participant is presented with cartoon representations of emotions. The methods in this cluster focus on feelings and sensations that UX may elicit. The second cluster concerns methods that exist for a specific use case, so a specific product or feature, such as the aesthetics scale [77]. In the third cluster, methods for extensive analysis are discussed. The ESM (Experience Sampling Method) [114] is mentioned as an example. They note that these types of methods are generally less suitable for projects in the industry, as they are often fast-paced and desire cost-effective methods. The fourth cluster focuses on qualitative evaluation, such as the DRM (Day Reconstruction Method) [63]. The fifth and final cluster concerns questionnaire-based methods, such as the Product Attachment Scale [89].

In their research Drouet and Bernhaupt identified three types of UX: momentary UX, episodic UX and cumulative UX. They did this to cover the time ranges present in the UX. Momentary UX was measured through a first impressions questionnaire. Then episodic UX was measured through some rating-questions that were asked after the user performed a task. These measure the subjective experience (e.g. emotion). The AttrakDiff questionnaire was applied to measure the cumulative UX.

Another distinction made between different forms of UX can be found in a study by Hasan, Morris and Proberts [52]. They categorise based on the way the UX problems are identified, in this case by users, evaluators or tools. User-based UXEMs record user's performance during an interaction with an interface. Users' preferences and satisfaction can also be taken into account. With evaluator-based UXEMs the evaluator is the person who identifies the UX problems and can also be found in the literature as usability inspection methods. Tool-based UXEMs involve software tools and other tools like models to identify UX problems. Hasan, Morris and Proberts state that user-based and evaluator-based methods are often used to evaluate the UX of websites.

Triangulation of UX evaluation methods is a topic found in the literature, stating that using multiple UXEMs can maximize the effectiveness of the UX design and evaluation process [104]. A common combination of methods is questionnaires and interviews [100]. Depending on the combinations of methods, the quality of the results may improve.

In the literature, many different UX evaluation methods exist. The 126 methods found in relevant papers in the MLR can be found in table 5 in alphabetical order. In Appendix A these UXEMs are elaborated upon. In the following subsections the six broad categories of UXEMs are elaborated upon. These categories serve as an abstract view on the essence of the evaluation methods, based on observed similarities between them.

Table 5: User experience evaluation methods found in the MLR

2DES	Geneva emotion wheel	Reaction checklists
Aesthetics scale	Goals, Operators, Methods and Selection (GOMS) rules analysis	Repertory Grid Technique (RGT)
Affect grid	Hedonic Utility Scale (HED/UT)	Resonance testing
After Scenario Questionnaire (ASQ)	Heuristic evaluation / guideline review	Self Assessment Manikin (SAM)
AttrakDiff	Human computer trust	Semiotic inspection method
Attrak-Work questionnaire	Immersion	Semi-structured experience interview
Audio narrative	Interview	Sensual evaluation instrument
Automated evaluation via software tool	Intrinsic Motivation Inventory (IMI)	Sentence completion
Canvas card sorting	iScale	ServUX questionnaire
Card sorting	IsoMetrics	Simplified pluralistic walkthrough
Click map / scroll map / heat map	Kansei engineering software	Simplified streamlined cognitive walkthrough
Co-discovery	laddering	Single Ease Question (SEQ)
Cognitive jogthrough	living lab method	Software metrics usability metrics
Cognitive task analysis	Long term diary study	Subjective Mental Effort Questionnaire (SMEQ)
Cognitive walkthrough	MAX	SUMI
Context-aware ESM	Mental effort	Survey / questionnaire
Contextual laddering	Mind map	System Usability Scale (SUS)
Day Reconstruction Method (DRM)	Multiple sorting method	Task environment analysis
Diary study	MUSiC performance measurement method	This-or-that
Differential Emotions Scale (DES)	Net Promotor Scale (NPS)	Timed ESM
Domain Specific Inspection (DSI)	Opinion mining / sentiment analysis	Tracking Real Time User Experience (TRUE)

EMO2	Outdoor Play Observation Scheme	TUMCAT
EmoScope	PAD	Usability & communicability evaluation method
Emotion cards / emocards / emofaces	Paired comparison / pairwise comparison	Usability guidelines
Emotion Sampling Device (ESD)	Participatory heuristic evaluation	Usability Metrics for User Experience
Experience clip	Pencil & paper	User Experience Questionnaire
Experience recollection method	Personas	User Model Checklist
Experience report	Perspective-based (usability) inspection	User testing - log analysis
Experience Sampling Method (ESM)	physiological signals / physiological UX evaluation / psychophysiological measurements / physiological arousal via electrodermal activity / facial EMG	User testing - performance measurement
Experiential contextual inquiry	Playability heuristics	User testing - question asking
Expert review / expert valuation	Positive and Negative Affect Scale (PANAS)	User testing - retrospective thinking aloud
Expert walkthrough (group based)	Presence questionnaire	User testing - thinking aloud / thinking out loud
Exploration test	Private camera conversation	User testing (extended usability testing)
Eye tracking	Product Attachment Scale	User workflow
Facereader	Product Emotion Measurement instrument (PrEmo)	User's feedback
Feeltrace	Product experience tracker	UTAUT
Field observation / field study / observation	Product personality assignment	UX Curve
Field study: 3E (Expressing Experiences and Emotions)	Product reviews	Valence method
Focus group	Product Semantic Analysis (PSA)	Web usability evaluation process
Fun toolkit	Property checklists	Website Analysis and Measurement Inventory (WAMMI)
Game Experience Questionnaire (GEQ)	Prototype evaluation	Workshops / probe interviews

Geneva appraisal question-naire	QSA GQM questionnaire	
---------------------------------	-----------------------	--

Expert evaluations The category of expert evaluations concerns any methods that center experts instead of users. Often these experts are user experience experts, but sometimes other experts such as domain experts are required. One subcategory in expert evaluations is heuristics. In the literature, heuristic evaluations are common under some different names and in different forms. Their common denominator is an evaluation based on any type of established guidelines. There are roughly two types of expert evaluation methods found in the literature. In the first type the experts look at the product from their expert point of view, where in the second type the experts simulates the actions of a user. Examples of methods in this category are heuristic evaluation, task environment analysis and user workflow. A list of expert evaluation methods can be found in Appendix A in table 12.

Field studies Any type of user experience evaluation method that concerns real-life contexts are classified as field study. For these methods it is important that the user is observed in their natural context. The method can either gather data during or directly after the interaction with the system, or some time after the interaction (e.g. at a pre-determined time during the day). Examples of methods in the field studies category are Day Reconstruction Method (DRM), experience report and immersion. A list of field study methods can be found in Appendix A in table 13.

Interviews Interviews and techniques that can be applied during interviews are grouped in this category. Methods in the interview category aim to ask the user questions, usually prepared in advance, for a specific UX related purpose. Interviews can take place as a singular user experience evaluation method or they can be used in addition to another method to gain a deeper understanding of the user. Examples of interview methods in this category are contextual laddering, exploration test and semi-structured experience interviews. A list of interview methods can be found in Appendix A in table 14.

Measurements Measurement methods are methods that require physical sensors to measure the responses of a user to a product or stimulus. Through sensors such as skin conductance or electromyography the physiological responses of a user can be measured. These methods can be combined with a self-reported method to find out what the user has experienced. It is possible to use tools for combining and analyzing data. The measurement method can be found in Appendix A in table 15.

Scales and questionnaires Scales and questionnaires are mentioned in the literature as abstract concepts and as concrete, established scales and questionnaires. In this context, a method is considered a scale or questionnaire when an established set of questions is presented to a user in some way. The scales that were found were self-reported scales

and thus subjective measures. Examples of methods in this category are the aesthetics scale, 2DES and SUMI. A list of scales and questionnaires can be found in Appendix A in table 16.

Software tools Any methods that require a software tool to be applied can be grouped under the *Software tool* category. Some methods were developed as software tools and are unique in their functionalities, other methods are a concept for which different software tools can exist. Methods that can be applied through software tools are heat maps, eye tracking and iScale. A list of software tool methods can be found in Appendix A in table 17.

Workshops In the category workshop methods are grouped that describe the concept of workshops or describe a method that can be applied during a workshop. In a workshop users are asked to do something, e.g. group concepts together, as opposed to interviews where questions are asked and no tasks are required. Examples of methods that can be applied during workshops are card sorting, prototype evaluation and this-or-that. A list of workshop methods can be found in Appendix A in table 18.

4.2 SQ2: What are the positive and negative aspects of each type of user experience evaluation method?

Each method brings its own unique set of challenges and benefits to the table. However, within categories of UXEMs there often are similarities. In this section these aspects are elaborated upon.

Expert evaluations In the methods in this category, the expert is centered. It can be challenging to find an expert on a certain topic. When the expert required is a UX expert, it is important to ensure this expert possesses enough UX knowledge to be seen as an expert as well as the correct skills to apply the desired method. Domain experts may need to be recruited, making the method more time consuming. Generally speaking, expert evaluations take less time to prepare as no users are required. An expert evaluation can help identify areas that need more focus and serve as a relatively cheap and quick way to keep shareholders involved and up to date. However, as an expert evaluates the system instead of a user, the problems that impact users may go unnoticed.

Heuristic are often quick and cheap. One expert can conduct a heuristic evaluation, but this can also be done in a group of experts. It is important that the person conducting the evaluation interprets the heuristics correctly. The heuristics that were found are established methods, but it is possible to develop new heuristics [102].

Field studies A large reason to conduct a type of field study is to observe users in their natural environment. Lab settings can have an influence on users' behaviours and a risk of observer bias where participants may not speak freely. However, the absence of an observer or researcher may increase the risk of less specific or lower quality answers in the case of self-reported methods, which is only found out after the experiment is

over. Methods where participants report their experiences directly during or after an interaction reduce the disturbance of memory effects where results may be less reliable when users have to think about what happened in the past. Another advantage of field studies is that the data gathered is rich, objective and detailed, though this makes it more time-consuming to analyse. Some methods do have a structured way to analyse the results, decreasing the amount of time it takes to perform an analysis.

Interviews Many types of interviews allow the researcher to ask the participant for elaboration or clarification on their answers. Interviews can uncover the real perceptions and needs of a user in a product. It can be challenging to ask the correct questions and to create a high quality interview protocol. The longer an interview is, the longer the analysis takes. Especially for longer interviews it is important that a skilled interviewer is present. Shorter, more structured interviews, provided the interview protocol is of sufficient quality, can be done by less experienced researchers. For interview methods that require the user to narrate their actions, the quality of the results depends on how comfortable the user is with narration. Interview methods can be used as a supplementary methods to other UXEMs to elicit more detailed information or to shine light on their thought processes.

Measurements Measurements are objective measures as sensors are used to measure the responses of a user. This objective measure may be combined with a self-reported method to gain context and learn more about the experience of the user. The sensors may limit the movement for the participant or make them uncomfortable.

Scales and questionnaires Scales and questionnaires are often quick and structured methods that can be applied by less experienced researchers. Scales should be validated to ensure they test the intended subjects. Many scales and questionnaires found in the literature are heavily researched and validated and are considered reliable methods. Questionnaires can also be developed for one specific purpose, but these are considered less reliable as they cannot be validated. Guidelines for creating questionnaires do exist [112]. Scales and questionnaires that are presented to a user are self-reported and thus subjective, leading to less reliable results. Many scales and questionnaires require functional prototypes or existing products, so these types of methods are not suitable for concepts and non-functional prototypes.

Software tools Methods in the software tools category require software to function. For more known methods many different types of software providers exist, making the software and thus the method more accessible. Some methods require one specific software product. Software can be inaccessible to some through a paywall and by using a specific software tool one may experience vendor lock-in, where it is highly cost ineffective and labor intensive to switch to a different software provider. Software tools collect objective data and allow for remote testing. An experienced UX researcher is often required to analyse the gathered data, as the data may be difficult to interpret. Software tools also require a functional prototype or fully functional product. Software

tools, especially ones used for automated UX, are relatively new and rapidly developing [3].

Workshops Methods in the workshops category often aim to visualise concepts and communicate these concepts to users. Visualising concepts can allow for easier communication with the user and sometimes even allow for non-verbal communication. Some workshop type methods require discussions to take place in a group setting. This can make the analysis a difficult and time intensive process. Other methods can be shorter and more to the point, as materials are easily created or provided and sessions can be shorter and done with individual users. These types of methods do not require an experienced researcher and can test products from an early development phase to fully functional products. Workshop methods also often allow for remote testing through synchronous sessions.

4.3 SQ3: Which of the found user experience evaluation methods are best suited for academic purposes and which are best suited for corporate purposes?

While the importance of UX is recognized more and more in the industry, the concept of user experience remains quite vague [7]. Alves, Valente and Nunes [7] conducted a study into user experience evaluation methods used in the industry and found that the following methods were almost always or always used: observation, think aloud, contextual interviews or inquiry, interviews, experience prototyping, task analysis, cognitive walkthrough, questionnaires, customer experience audit and KPI (Key Performance Indicator). Methods that were never used were Kano analysis, semantic differential, ESM (Experience Sampling Method), eye tracking, behavioural maps, critical incident technique, shadowing, desirability testing and, overlapping with almost always used, KPI and customer experience audit. They state that informal, low cost methods are preferred in the industry and working prototypes are favoured. They also found that evaluations often happen at multiple phases within a project and that various methods are used for this. Finally, Alves, Valente and Nunes found that evaluations can be constrained by the experience or occupation of the evaluators.

Ardito et al. [8] confirmed in their study that UX and usability is still not given enough priority in companies, with these concepts being either neglected or not being properly considered. Through interviews and focus groups they found that it was thought that involving end users is a waste of time, as the users cannot explain their needs or expectations. Ardito et al. call for public organizations to actively consider UX and usability and state that *"...it is [the] responsibility of academics to translate scientific articles, which formally describe evaluation methods, into something that makes sense for companies and is ready to be applied"*.

Bang, Kanstrup, Kjems and Stage conducted a research study which describes how UX methods can be introduced in an IT organization and how the industry can prioritize UX decisions [12]. They found that academic evaluation methods can use words that sounds too academic and often require some modifications to be able to use them in industry settings. They state that scientific papers often have different goals and purposes for evaluation methods and thus need to be altered to suit the more practical needs

of companies. Lachner et al. found in their study regarding quantified UX (QUX) that the understanding of UX principles and the product development processes differ a lot between companies [76]. The processes may be less structured in start ups, while their view of UX might be more holistic. Their proposed QUX serves as a first step towards establishing UX as a unified measurement approach.

Väänänen-Vainio-Mattila, Roto and Hassenzahl created a list of requirements for practical UX evaluation methods [121]. They state that for UXEMs to be used in a practical setting, they should be:

- Valid, reliable and repeatable to ensure UX can also be managed in a large company;
- Fast, lightweight and cost-effective to suit the fast-paced iterative development cycles;
- Low expertise level required so that the method can be easily applied without extensive training;
- Applicable for various types of products so products can be compared and trends can be monitored;
- Applicable for concept ideas, prototypes and products to suit the different stages of the UX development process;
- Suitable for different target user groups to ensure a fair outcome;
- Suitable for different product lifecycle phases for improving taking into use and repurchasing UX;
- Producing comparable output (quantitative and qualitative) for iterative improvement;
- Useful for different in-house stakeholders to allow for easier information sharing between departments.

Väänänen-Vainio-Mattila, Roto and Hassenzahl state that not one method can fulfill all of these criteria at the same time. Because of this, it can be difficult to decide what method to use.

It seems that in the industry, preference is given to low cost, informal methods that can be applied to working prototypes. If an academic evaluation method is to be adopted in a company, changes may need to be made to suit their more practical needs. Based on this, it can be concluded that methods that yield quantitative results are generally more suitable to apply in practice. Of course, this does depend on the purpose of the user experience evaluation and on the time and funds that are allocated to a project.

4.4 SQ4: What decision processes for user experience evaluation methods currently exist in academic and corporate settings?

When searching for existing decision methods regarding user experience evaluation methods, few relevant papers come up. A total of six sources were found in the MLR that concerned a description of the UXEM decision making process.

Decision making Darin, Coelho and Borges performed a systematic snowballing procedure to compile a list of instruments, meant to assist researchers and practitioners

in their decision making process [27]. They provide a classification of UX instruments based on nine application domains and created a detailed list with more information. This list may be useful to researchers, but misses some crucial information to practitioners such as strengths and weaknesses of instruments and constraints such as time and funds.

Dhouib et al. [34] reviewed the literature and found the main factors that can affect the selection of usability evaluation methods, specifically regarding interactive adaptive systems. They use three common usability evaluation methods (heuristic evaluation, usability test and cognitive walkthrough) and consider three groups of criteria (situational factors, characteristics of stakeholders and adaptivity aspects). Situational factors include the stage in the development life cycle, temporal and financial resources, the style of evaluation and the type of data. Characteristics of stakeholders include the number of users and evaluators involved, the availability of direct access to users and the level of expertise of evaluators. Finally, the adaptivity aspects include reusability adaptation rules and intrusiveness of adaptivity. The adaptivity aspects may be too specifically intended for interactive adaptive systems, but the other two groups seem more generalizable for other UXEMs.

Decision tools Melo and Jorge [88] present a mechanism to help UX decision and evaluation method identification. They use multiple criteria decision-making tools. The proposed mechanism should be used when choosing a method that should fit certain requirements or capabilities. To determine whether the method would fit the situation, they use the following criteria: depth (in-depth or not, scored on 5 point Likert scale), documentation level (how well documented is a method, scored on 5 point Likert scale), structure (degree of formality, scored by very informal, informal, formal, very formal), time (30 seconds to several years), expertise required (scored by novice, familiar, knowledgeable, expert) and phase (design concept, mockup/paper, functional prototype, functioning product). This tool can be applied in any stage of the design and evaluation process, and the methods used in this paper were heuristic evaluation, cognitive walkthrough, UX curve, eye tracking, card sorting and contextual inquiry.

Fischer, Strenge and Nebe [43] describe the process of developing a tool to aid in method selection. They state that in the literature, usability is often seen as a one-dimensional construct. They argue that the outcome of a usability method selection tool should be a set of methods that covers multiple aspects of a product's usability. The theoretical concept for the tool described in the paper is based on the ISO/TR 16982:2002, a standard aimed at project managers. Some of the measurements in the tool, such as number of hours required, are community-based. Though this article is from 2013, no updates were found and the concept seems to remain theoretical.

User type Filippi presents an evaluation of PERSEL [40]. This tool selects the ideal personality of users for UX redesign activities. Users are first asked to answer a questionnaire on which PERSEL can base the decision. When selecting user personalities, PERSEL looks at a total of UX items such as usefulness, usability, aesthetics and emotions. PERSEL is focused specifically on the redesign phase, as it is generally more

structured than the design phase. In their paper, Oyugi, Abdelnour-Nocera and Clemensen also emphasize the need to consider user type in UX evaluation methods [93].

The tools discussed in these subsections are concepts, models or software tools. The concepts and models can be found in research papers or websites, where the software tools were created to be used on desktop computers. Decision processes can therefore take place on paper, on a website or on a desktop computer.

5 Results: Contextual inquiry

In the previous section, the literature review has been conducted based on the four research questions that were to be answered through the literature. In this section the contextual inquiry is elaborated upon. The contextual inquiry aims to inform the answers to sub research questions SQ1, SQ2 and SQ4 and helps formulate requirements for SQ5. The contextual inquiry is part of process A: *data collection*. This section is part of process B: *data analysis*. This is shown in figure 8.

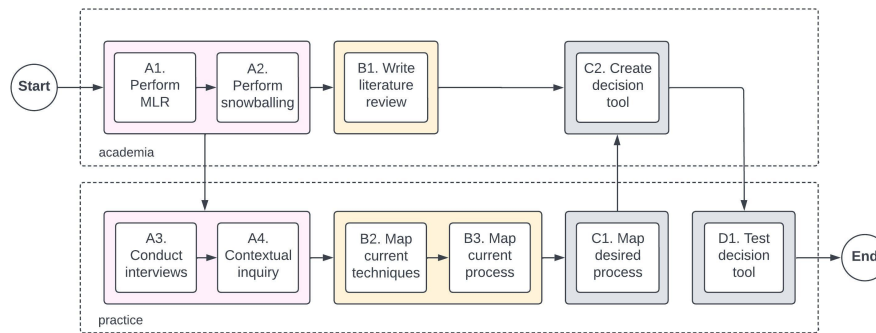


Fig. 8: Process A and B: data collection and analysis

It is important to note that there are some risks to this method. The researcher applying the contextual inquiry method may bring their own bias. This risk can be reduced by going into the research with an open mind, to not make assumptions and to treat every new piece of information with the same level of care and importance. Observer bias can occur when a user changes the way they operate when someone is watching them. It might cause stress to have someone watching over your shoulder. The risk of observer bias can be partially reduced by stressing the importance of carrying on as they normally would and by having open-ended conversations where the users fill in the blanks for the researcher. It is also a risk that users skip over the details of their work by going into what Salazar [109] calls interview mode. This can be reduced by reminding the user that the small details are important to know. Finally, users may air their grievances about the current way of working. While this is understandable, this is not the goal of the sessions. The researcher should steer the discussion back on topic if this occurs.

5.1 Current situation

To investigate the current way of working in UX teams across the company, employees from different locations were invited to an informal discussion, either in real life or through video calling. In this section the insights from those discussions are reported.

To elaborate on this contextual inquiry, semi-structured interviews are to take place with UX team members from the different locations. With this information, the current situation regarding present UX evaluation methods and decision method could be mapped in detail. In Appendix B, the informed consent can be found and in Appendix C the interview protocol can be found.

The locations are selected based on availability of UX team members and presence of UX teams. Some locations have many open projects and thus the UX teams do not have time to participate in interviews.

Location A Location A is the location where most of the contextual inquiry takes place, as this is where the author was present. The UX team in location A falls under the Content and Creation branch. When a new project starts, the team lead chooses a team member based on skill set and availability. Usually there is one team member per project, sometimes there are two. This team member chooses the UX evaluation method they see fit, so the team largely relies on the existing knowledge within their team. Every week, the UX team meets to discuss any problems they may have with the other team members, to ask for help or to update them on their project.

The process at location A is described in figure 9. This figure shows the current decision process at this location, from the kick-off to the start of the visual design. As location A is where the contextual inquiry takes place, the information on this location is more detailed than other locations.

Location B At this location, the UX team is under the Technology branch instead of Content and Creation. This is because of the technical background of the teams, as they have experience with implementation of UX solutions.

At location B, the UX team chooses a suitable UX method based on the GOTIK method. This method can be used for project management and consists of five components (translated from Dutch): money, organization, time, information and quality [1]. The method is also known under different acronyms but usually contains at least the five components mentioned, sometimes with one added depending on the type of organization. The GOTIK method assists in deciding which way of working is most suitable for the specific project. The team identifies the most important aspect of the five components and suggests a work flow based on this. Following this, a project that should be completed quickly might cost more, and a project that has a higher budget may be of higher quality.

In their Sharepoint folder, the team started to keep a list of the research methods they use and any important documentation that could be necessary. This folder is not currently in use, but does offer insights into the different methods that are used.

Location C Location C sometimes collaborates with location B on certain projects, but this does not happen very often. At location C the roles of UX researcher and UX designer are clearly separated, so UX researchers receive projects for testing from designers. To determine an evaluation method, the researcher asks the designer what they want to test in the prototype and bases the decision off the answer from the designer

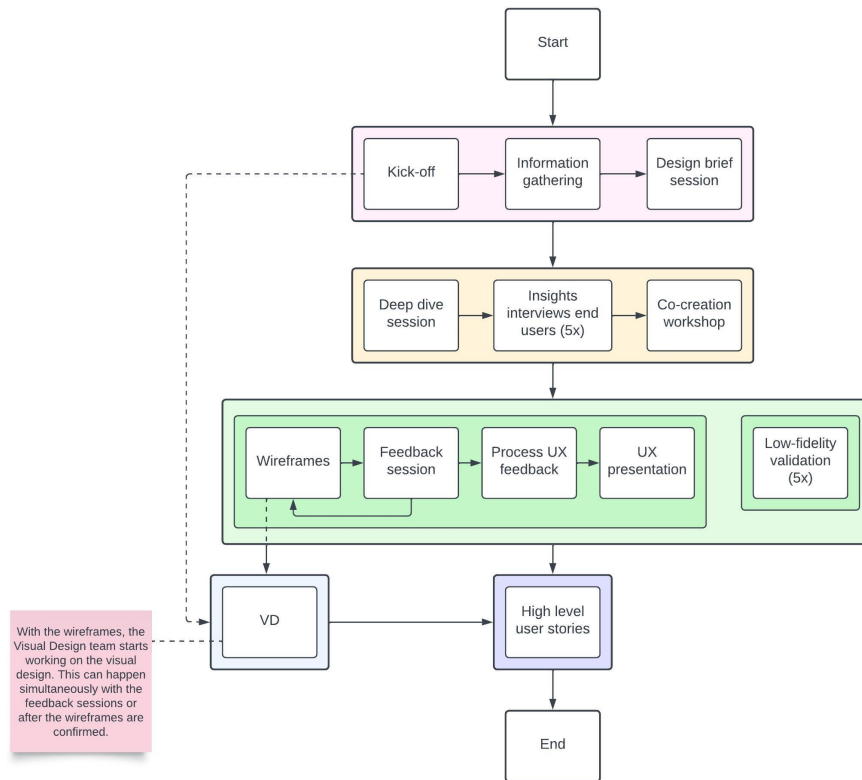


Fig. 9: UX team process at location A

and their own expert knowledge. There is no known UX evaluation methods database to collect or share information between team members.

Location D At location D members of the UX team often work on long-term projects with the same clients. These clients are often larger non-IT companies. Within the long-term projects, shorter projects exist. To determine what UX evaluation method should be used, some UX team members use a table by Rohrer [108] where methods are sorted based on their development phase (strategize, execute and assess). This table is not used as a golden rule, but more as a guideline for inspiration. There is no known UX evaluation methods database to collect or share information between team members.

5.2 Current set of evaluation methods

In table 6, the outcome of the preliminary interviews is shown. The types of research method are methods mentioned by the UX team members. A check mark means the knowledge for a specific UX evaluation method is present, but this does not necessarily mean that the evaluation method is used frequently. During the conversations with team members, it became clear not all used terms for the types of evaluation methods were uniform. This has been corrected in the table.

Table 6: Set of UX evaluation methods present

Type of evaluation method	Location A	Location B	Location C	Location D
Analytics tool	✓	✓	✓	✓
Concept testing & validation	✓	✓	✓	✓
Desk research	✓	✓	✓	✓
Expert reviews	✓	✓	✓	✓
Eye tracking	✗	✓	✗	✗
High fidelity (clickable) prototype(s)	✓	✓	✓	✓
Interviews	✓	✓	✓	✓
Neuro measurement	✗	✓	✗	✗
Surveys	✓	✓	✓	✓
Usability testing	✓	✓	✓	✓
UX audits	✓	✓	✓	✓
Wireframe prototype(s)	✓	✓	✓	✓

The most notable difference in the table is seen in *eye tracking* and *neuro measurement*. Location B possesses knowledge to apply these techniques, where other locations do not typically reach for them.

5.3 Requirements

During the contextual inquiry, requirements for the decision tool were gathered. From the information gathered during the contextual inquiry and interviews, the following user stories could be created:

- US1: As a user, I want a decision support tool for user experience evaluation methods, so that I can be sure I made the right choice.
- US2: As a user, I want to share the results with my colleagues or clients so that I can easily explain my reasoning.
- US3: As a user, I want to be able to complete the decision process quickly, so that it does not add to my busy schedule.
- US4: As a manager, I want the information to be up to date so that the customers receive the most accurate information.
- US5: As a user, I want to connect with my colleagues so that we can share information and learn from each other.
- US6: As a user, I want to expand my knowledge on UX evaluation methods so that my expertise keeps growing.
- US7: As a user, I want to save my outcome so that I can reflect on them later.
- US8: As a user, I want to customize my profile so that I can control the information my colleagues can see.

Based on these user stories, the requirements for the system were created. These can be found in table 7.

Table 7: Requirements for the system

Requirement	User story
Decision process support	US1
Share button to email and messaging apps	US2
Export to PDF	US2
Recap the answers of the decision process before submitting	US3
Skip button	US3
Option to add new methods to the database	SQ4
Contact information to reach information manager	US4
Contact option to quickly reach information manager	US4
See colleagues and their expertise	US5
Introduce lesser known methods	US6
Create a profile	US7 & US8
Profile customization	US8

6 Results: Interviews

The research questions that could be answered through interviews and contextual inquiry are SQ1, SQ2, SQ4 and SQ5. In the previous section the contextual inquiry is described. The interviews are discussed in this section. The combined answers to these questions were used to create the decision tool. The interviews were conducted in process A: *data collection* and described in this section as a part of process B: *data analysis* as shown in figure 10.

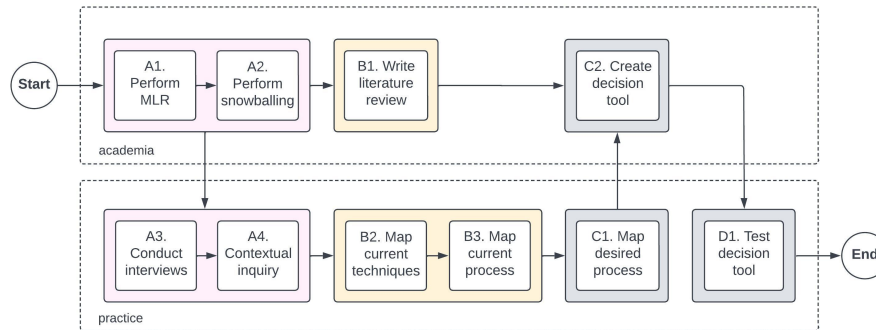


Fig. 10: Process A and B: data collection and analysis

Interviews are prone to memory problems. As human memory is flawed, people might misremember information or events or even not remember them at all. It is also important to note that participants may leave out details, as they do not know what information is important to the researcher. The interviewer might need to ask further questions to elicit the desired information. Finally, some people are uncomfortable with sharing a lot of details with a stranger or may be shy.

Semi-structured interviews were conducted with two members of four different UX teams across four campuses (Location A, B, C and D). The campuses and the team members were chosen based on convenience sampling with location A as a basis. The campus proximity was an important contributor to the sampling method, as it is likely that campuses that are closer together have more face-to-face interaction. All campuses selected for the interviews were in the Netherlands and all interviewees were Dutch. In this section the insights from the eight interviews and from contextual inquiry are elaborated upon.

The interview protocol can be found in Appendix C. Before the interviews started, participants were presented with an informed consent form (Appendix B).

6.1 Way of working

Location A At location A projects are typically assigned through the project manager and the sales department. They are the first to know about any new projects coming up.

The projects are passed on to the team lead, who divides the projects based on time available, team capacity, seniority and personal strengths. Sometimes two colleagues are put on the same project, where one has the final responsibility and the other is available for questions.

At location A, the members of the team have created a sheet describing the steps one would take in an ideal situation. However, almost no projects are truly an ideal situation and thus changes and shifts must almost always be made. Team members are free to determine their own process. Choices are based on the amount of time available. The time available is a set number of hours that are allocated to this part of the project by the sales department and the project manager.

"Usually one can do less research than is desired. It is always a dance between contents and available hours."

Choosing an evaluation method often happens subconsciously, one participant states. As you learn more about a project, the ideas start to shape in your mind. There are many similarities between projects, as the company usually designs web pages, so the concept is largely the same. The team is open to new methods, but one participant states that everyone uses the same few methods for the majority of the projects. It would be time consuming to switch to a different method that you may know less about. Finding participants can also be a challenge.

The evaluation process at location A is often summative, due to time constraints, but for long-term projects formative evaluations are becoming more common. This is because both parties would like the collaboration to last a long time.

The participants from location A think that if a colleague from the same team would work on their project, the outcome would largely be the same. The projects are comparable and the general outline of the evaluation process is described. All team members use a similar set of preferred UXEMs, with some differences based on personal preference and skill set. One participant described the way of working within the team as *"everyone is on their own island"*, meaning that team members only have a general idea of what their colleagues are working on.

When asked what the participants think colleagues from other locations would do, they stated that they were not sure. One of them expected the results to be largely the same. The other participant thought there would be significant differences. The participants were unaware of the exact way of working at other locations.

"I think it is very unfortunate how user experience evaluation research is the first to be eliminated in a time crunch."

Location B At location B, projects are assigned based on complexity of the problem, as complex problems often require more experience. The available number of hours for a project is taken into account and checked against the availability of UX team members. It is also important to note personal expertise and preference. Finally, personal factors can also be taken into account, so if the office of a new client is close to where a UX team member lives, this can also influence the decision.

Senior employees often determine what should be done for their own projects, where junior employees may do this in consultation with a senior employee. A schedule is made by the project managers, and the team members write down the steps and the number of hours spent. Evaluation methods are chosen based on what needs to be done and what needs to be tested. Smaller flows or usability problems can require smaller, shorter evaluation methods. Larger changes or flows may require a more extensive evaluation method. It is also important to know whether qualitative or quantitative data is required. In the case of a new product or prototype, it can take up until the first functional prototype to decide on an evaluation method.

At location B, both formative and summative evaluations take place. At the end of a design process, an evaluation is performed. Based on the results, it is decided whether the project is done or should be continued. It also depends on the stage of the development process the evaluation takes place at.

Both participants think their colleagues would work in a similar way to them. One participant stated that they sometimes assist their colleagues in choosing a method, so the set of methods is more or less the same. The other participant stated that they thought the quality of the research and the experience of the researcher would make a difference, but that the methods would be largely the same.

"It is also about the research style, such as guiding questions versus asking for more clarification."

Regarding colleagues from different locations, one participant stated that their answer is the same as to the other question. They expect most UX team members, also from other locations, use a similar set of a few well-known methods. The other participant did not know, as they had not interacted with colleagues from other locations.

Both participants made a comment about the degree of academic language present in the list of methods.

"I think if we were really going to implement methods from this list, methods should be shorter and to the point."

Location C At location C, the assignment of projects depends on one's role. There is a distinction between UX researchers and UX team members, where UX researchers are flown in just for the tests and the UX team members are involved with a larger part of the project. The UX researcher can be contacted by one of the team members to test a prototype. Within the UX team, the project assignment is done by the team lead in close contact with the team members. Based on time available, seniority and availability, projects can be assigned.

Team members at location C are free to choose their preferred way of working in a project and are thus free to decide what evaluation method to use. This decision can be made individually or as a team. There is no structured process, though one participant stated that they would like to see a more structured process. The decision also depends on the time available, the budget, the type of customer and the type of product. When there are multiple options, sometimes a customer is asked to choose. They almost al-

ways choose the less expensive option.

"When we present the customer with options, it is a shame that the customer often chooses the cheaper option."

Evaluations are often summative at location C, but there is a mix of both. One participant states that they often know what they want to design through prior research, so they finish an iteration by evaluating.

When asked how they thought their colleagues would approach one of their projects, the participants from location C stated that everyone has a preferred way of working, but the expectation is that the final result will be very similar. All members apply a similar set of methods. When a colleague from another location would take over, though, they stated that they had no idea. It depends on the experience of the person, and on the type of projects they are used to. Both participants state that they do not know much about the way of working at the other locations.

"My choice in methods is also based on the fact that I do not have much experience yet. It is more difficult to tell a customer why we are applying a certain method."

Location D At location D, projects are typically assigned by the team lead. The customer contacts a business consultant, and the business consultant contacts the team lead. The choice is mainly made based on experience, sometimes the team decides together. Then, the UX team member decides the steps that should be taken in a project, sometimes together with the customer. The method(s) chosen also depends on the personal preference of the team member. Some people are more familiar with scientific scales, others with visual design. Within the team at location D, the team members ask for peer feedback and share projects before they go live. A UX method is chosen based on the deadline, so the time available, budget, capacity. *"The golden triangle"*. The development phase also plays a part.

The evaluations that take place are largely summative, but sometimes formative evaluations take place as a main focus within the team is to always keep improving. For shorter projects the evaluations are often summative.

"We are all trying to answer the same questions."

When asked how they think a colleague would handle one of their projects, one participant stated that they were unsure, but that the final result would probably be the same. As there is no decision tree, all choices are made largely based on experience. The other participant made a similar statement, but did not think the process would be handled in the same way as there is no defined way of working. Both participants did not know the way of working at other locations. One participant did state that they suspect that other methods are used at different locations, as the locations have a wide range of different types of customers.

"The literature often does not rhyme with the reality of business."

6.2 Methods

Table 8: What methods were known and used by the participants?

Card sorting	Prototype evaluation
Click map / scroll map / heat map	Semi-structured experience interviews
Co-discovery	Software metrics / usability metrics
Expert review	Survey / questionnaire
Expert walkthrough (group based)	Usability guidelines
Eye tracking	User testing
Field observation / field study / observation	User testing - performance measurement
Focus group	User testing - question asking
Heuristic evaluation / guideline review	User testing - remote testing
Interview	User testing - thinking aloud
Mind map	User workflow
Personas	User's feedback
Private camera conversation	Workshops & probe interviews
Product reviews	

A total of 126 methods were presented to the participants. They indicated whether they knew the method and whether they have used it at the company. 27 methods were known and used by at least four out of eight participants. The other methods were either unknown, unused or used by three or less participants. Some methods had clear reasons for not using them (e.g. diary studies take very long and are difficult to analyse), others were simply less known as they are mostly used in academia.

Looking back at table 6, the interviews confirm that these methods are present. However, some methods used in the company are not mentioned by name in the MLR list, like desk research and UX audits. From the interviews, it became apparent that these terms often do not mean one single UXEM. Like usability testing, these terms are used as umbrella terms for a range of methods. An UX audit can consist of heuristic evaluations, interviews, usability or user tests, desk research or reviewing analytics. Knowing this, the following methods were not explicitly mentioned but implied:

- Click map / heat map / scroll map under analytics tool;
- Heuristic evaluation / guideline review under UX audit and usability testing;
- Product reviews under desk research;
- Semi-structured experience interviews under interviews;
- Usability guidelines under UX audits;
- User testing, including performance measurement, question asking, remote testing and thinking aloud under usability testing;
- Workshops and probe interviews under usability testing.

Methods that were not found during the contextual inquiry but were used:

- Card sorting;

- Co-discovery;
- Expert walkthrough (group based);
- Field observation / field study / observation;
- Focus group;
- Mind map;
- Personas;
- Private camera conversation.

When looking at the UXEM categories that were previously established, expert evaluation, interviews, software tools and workshops are the most frequently used types of methods. Field studies and scales and questionnaires were used less.

7 Implementation

In this section, a prototype for an interactive decision tool is proposed based on the MLR performed in section 4, contextual inquiry as described in section 5 and the interviews described in chapter 6. The prototype is created using Figma. This section describes process *C: implementation* as shown in figure 11.

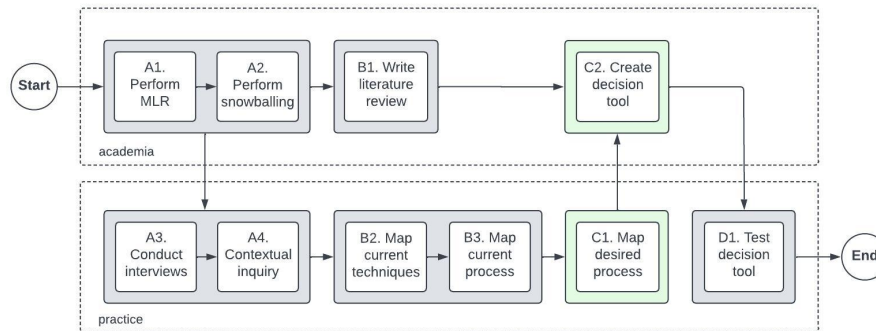


Fig. 11: Process C: implementation

The interactive decision tool aims to aid the user in the process of deciding what UXEM to use. It should provide the user with the most fitting UXEM(s) for a certain situation. Through a series of questions, the tool determines what solution the existing situation requires. Figure 12 shows a model of the general process of using the tool. The database can be edited by the team leads to add more evaluation methods or to update information. Users can edit their own profile. The full set of user experience evaluation methods can be found in Appendix A.

7.1 Mobile application

The decision tool takes the shape of a mobile application to underline the importance of a quick decision process. The tool should not make the decision process longer, which is challenging as the interviews indicate that there are no decision processes for choosing a UXEM. A mobile application was chosen for this so that UX team members can refer to the tool even when they are not at their work station (e.g. team meetings where information is shared or updating each other on current projects at the coffee machine) and the entire process should not take longer than a few minutes, possibly shorter after the app becomes more familiar.

The data used in the application can be managed by one information manager, who is familiar with the data set. If team members come across a new method that is not in the system yet, it can be added by the information manager. It is their responsibility to keep the data neatly organized and up-to-date.

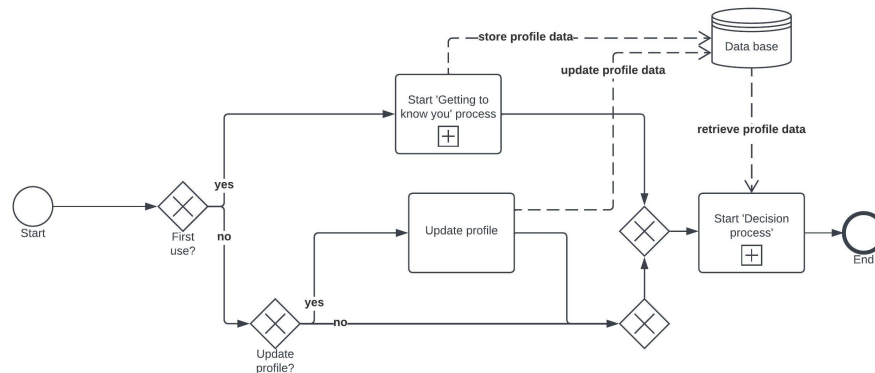


Fig. 12: Overview of the decision process

7.2 Requirements: MoSCoW

In this section, the requirements of the decision tool are discussed based on the MoSCoW prioritization method. Requirements in the *Must have* section are required for the tool to work. *Should have* requirements are high-priority items that should be included, if possible. *Could have* requirements are desirable features, but not necessary. *Will not have* requirements will not be implemented right now, but might be an option for the future. This analysis is based on the user stories and requirements presented in chapter 5 in table 7.

Must have The tool must include the create-a-profile process and the decision process. The user must be able to save and edit their profile. During the decision process, the user must be able to go back to correct a mistake.

Should have The tool should have a share feature, so the results of the decision process can be shared with colleagues or customers. The decision process should include a skip button, in case information is unknown. The user should be able to check their answers before they end the decision process so they would not have to start over to correct it.

Could have The system could provide contact information for the information manager, so any new methods can be sent to them immediately. The system could contain a way for the information manager to update the data used in the app.

Will not have The system will not include a social feature, where users are able to see who is familiar with certain methods. The system will not contain an introduction for lesser known but relevant methods. The system will not contain any further customization of the user profile, such as profile pictures or an 'about me' section.

The prototype of the tool contains all requirements from the *must have* and *should have* categories.

7.3 Content

Table 9 shows the questions that the tool asks the user upon first use and the possible answers the user can enter. These questions aim to elicit basic information about the user in the context of the UXEM decision process. FQ1 adds familiarity, the application can use the name to 'talk' to the user. The name is also important for the profile, especially for the information sharing and social aspect of the tool (although that is considered a *will not have* in the MoSCoW analysis above). FQ2 aims to find out the experience level of the user. FQ3 is intended to help create a list of known methods with which the user is very familiar. FQ4 aims to elicit information on methods that the user has heard of, but has not used (much). FQ5 is intended to confirm the choices made in FQ3 and FQ4.

Table 9: Questions upon first use

ID	Question	Possible answer
FQ1	What is your name?	Open ended
FQ2	What is your position at the company?	Choose from: junior/medior/senior + function title (open ended)
FQ3	Which of the following methods do you use often?	Choose from list of UXEMs
FQ4	Which of the following methods have you heard of, but not used often?	Choose from list of UXEMs
FQ5	Does that mean you do not currently possess knowledge regarding these methods?	Choose from: yes/no, edit my answers

Table 10 shows the questions that the tool asks the user when the decision process has started and the possible answers the user might enter. These questions are based on the matrix items in table 11, which are based on the previously conducted literature review and interviews. Each question focuses on a different aspect of the matrix. Q1 in table 10 aims to find out what type of product should be evaluated: a concept, non-functional prototype, functional prototype or a fully functional product? Some methods are unsuitable for one or more of these development phases. Q2 concerns the task that should be tested. Is it one task, or multiple tasks over a longer period of time? With Q3 a time frame can be established. How long does the user have to complete this project? If they have more time, they might be able to try out a new UXEM. Q4 concerns the type of outcome, with an option to choose 'no preference' if it is unclear. Q5 gives the user the chance to choose to work remotely or on location. This might be subject to personal preference or in accordance with the customer. Q6 is important to consider, as this may change depending on the type of customer.

During the MLR, several aspects of UX evaluation methods became apparent. From the interviews it became clear which aspects are most relevant to the decision process

Table 10: Questions from the decision process

ID	Question	Possible answer
Q1	Where are you in the development phase?	Choose from: concept/non-functional prototype/functional prototype/fully functional product
Q2	What needs testing?	Choose from: snapshot/one task/set of tasks/long-term
Q3	How much time is available to you?	Choose from: short/medium/long
Q4	Do you need qualitative data, quantitative data or both?	Choose from: qualitative data/quantitative data/both/no preference
Q5	Do you need to work remotely?	Choose from: yes/no preference
Q6	Do you want to observe users in their natural context?	Choose from: yes/no/no preference

as it exists in its current form. Based on the results from the interviews and from the MLR, a set of aspects was chosen as matrix items. These items are scored based on the answers to the questions in 10 and the familiarity variable based on the user profile, generated through the answers to the questions in table 9.

Table 11 shows which matrix items are influenced by the answers to the questions in table 10. Based on the answers to M1 and M2, evaluation methods can be eliminated. A score is assigned based on the other methods. Within each matrix item, a score of 3 points is given to the methods that fit the answer to the question the best. A score of 1 point is awarded to the methods that do not quite fit the answers. A score of 0 points is awarded to the methods that do not fit the answers.

Table 11: Influence of questions on matrix items

Question	Matrix item	Score
Q1	M1: Development phase	Eliminate non-fitting methods.
Q2	M2: To be tested	Eliminate non-fitting methods.
Q3	M3: How long does the method take?	3 points to best fitting; 1 point to medium suitable, 0 point to not suitable.
Q3	M4: Effort from researcher required	3 points to best fitting; 1 point to medium suitable, 0 point to not suitable.
Q4	M5: Qualitative or quantitative?	3 points to best fitting; 1 point to medium suitable, 0 point to not suitable.
Q5	M6: Remote possible?	3 points to best fitting; 1 point to medium suitable, 0 point to not suitable.
Q6	M7: Location	3 points to best fitting; 1 point to medium suitable, 0 point to not suitable.

7.4 Prototype

Based on the information above and the requirements that SQ5 yielded, a clickable prototype was created.

First use When the user first opens the application, they are presented with a screen asking them for their name. After entering their name and pressing the button to continue, the next question is presented.

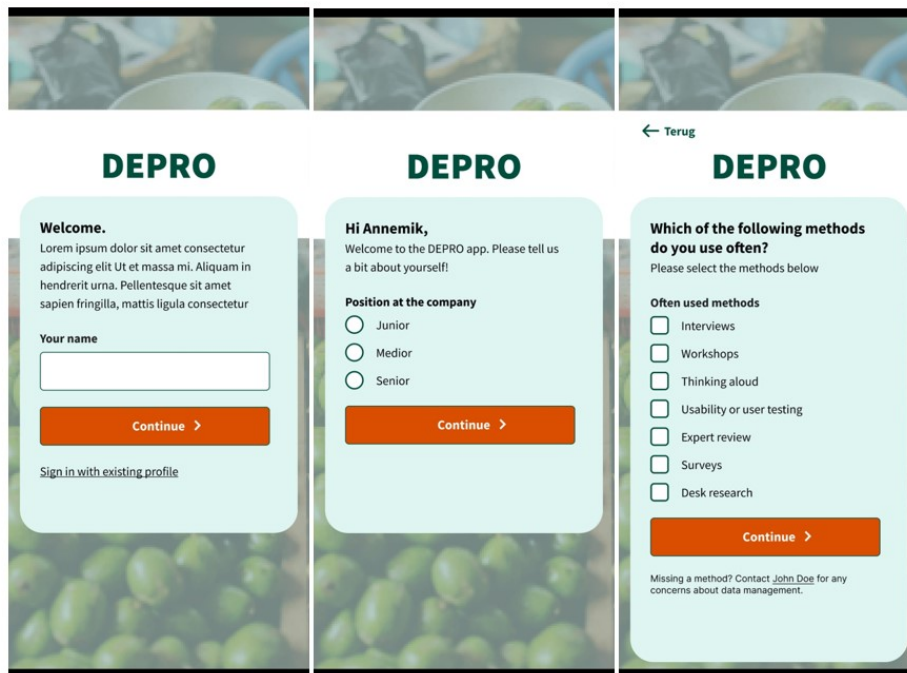


Fig. 13: First use process: introduction, company role and familiar methods

The user is asked to enter their position at the company. In the next question the system shows a list of UXEMs to choose from. The user can press one or multiple methods and scroll down for more. Once they have selected the methods they are familiar with, they move to the next question. Here, the user can select which methods are known to them, but not familiar. Upon selecting one or several methods, the user is presented with a summary of their answers to confirm their entries. After clicking the button to continue, a user profile is created. This profile can be edited to reflect recently gained knowledge or to correct errors. The user is then lead to the home page of the application, where they can see their previous projects or choose to start the decision process.

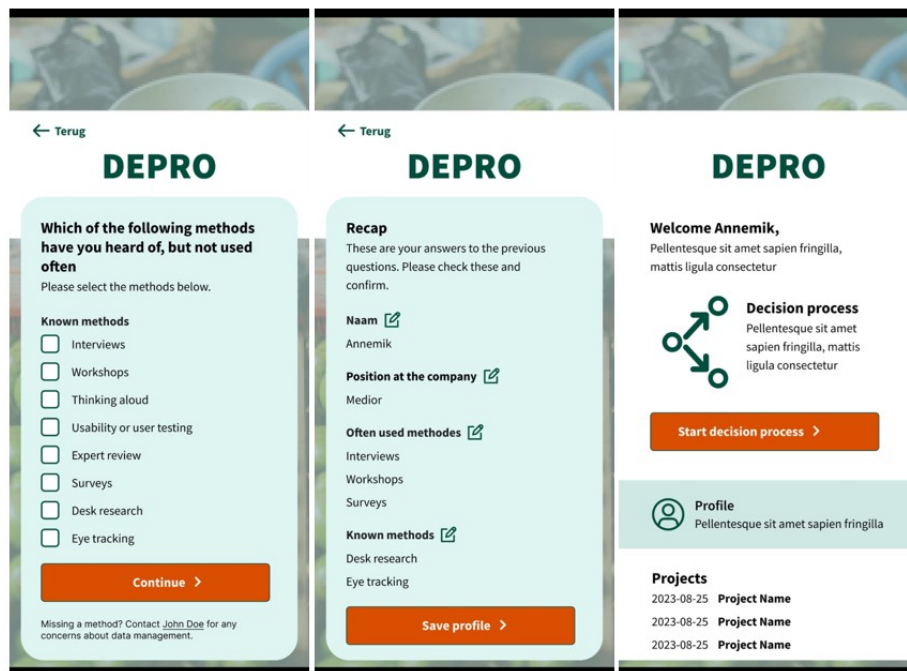


Fig. 14: First use process: known methods, recap and dashboard

Decision process Upon opening the application, the user finds themselves on the home page of the app. Here it is possible to view past projects or to start a new decision process.

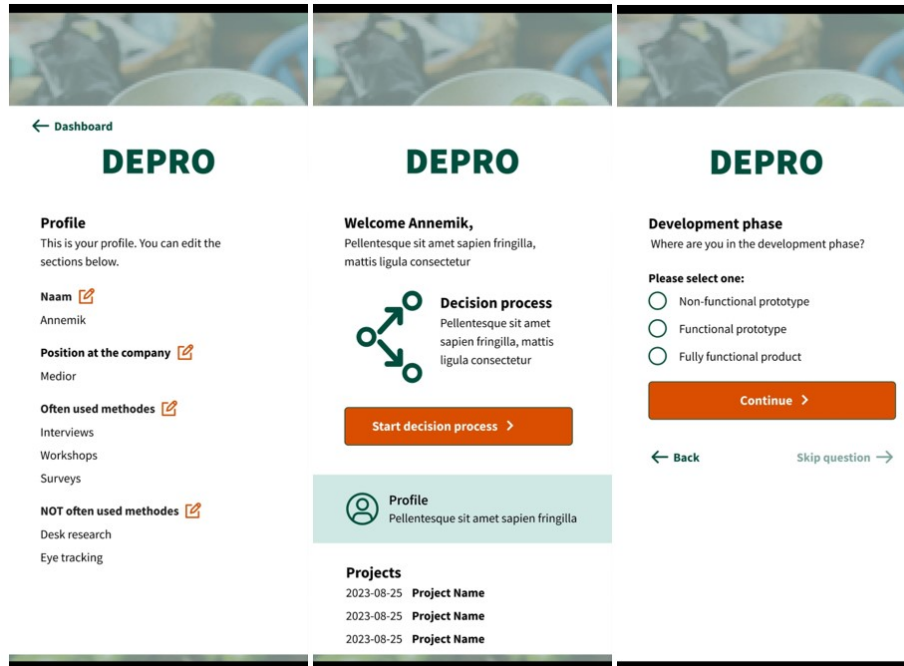
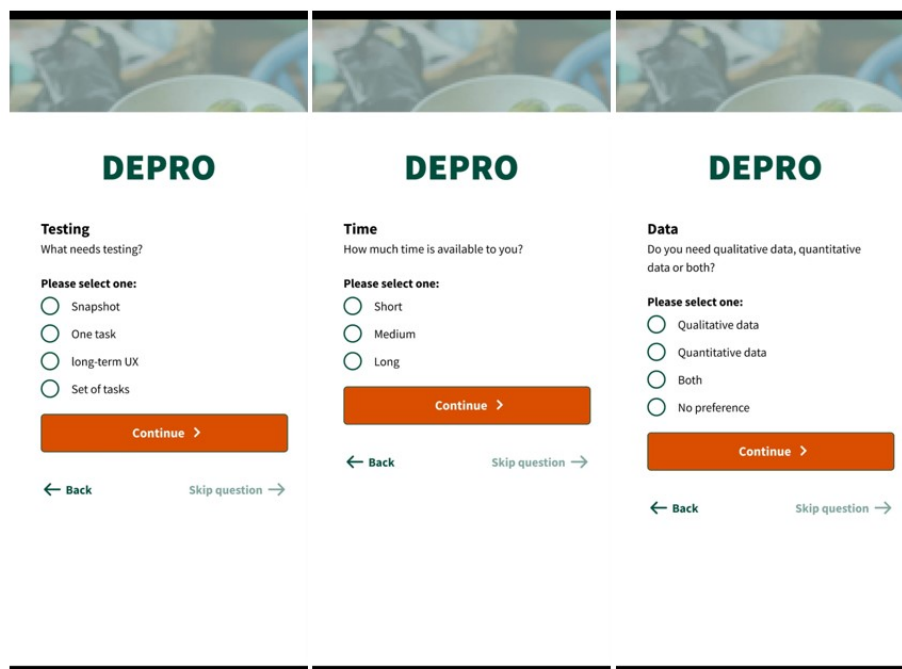


Fig. 15: Decision process: profile overview, dashboard and first question of decision process

The system leads the user through a series of questions discussed in previous sections. The user can navigate back to a previous question to correct a mistake or skip a question if they do not possess sufficient information to answer the question. After answering all questions, the user is shown a recap to correct any mistakes. They can then click the button to continue and are presented with an overview of the most suitable methods for this project. The system shows the suitability of the processes in percentages based on the matrix items in the sections above. The user can save this project to their profile for easy access in the future. After saving, they are returned to the dashboard.



The image displays three sequential mobile app screens for the DEPRO decision process. Each screen features a header with the word "DEPRO" in bold green text. The background of each screen is a blurred image of a person using a laptop at a table.

- Screen 1: Testing**
Title: **Testing**
Question: What needs testing?
Please select one:
 Snapshot
 One task
 long-term UX
 Set of tasks
A large orange "Continue >" button is positioned below the options. At the bottom, there are "← Back" and "Skip question →" links.
- Screen 2: Time**
Title: **Time**
Question: How much time is available to you?
Please select one:
 Short
 Medium
 Long
A large orange "Continue >" button is positioned below the options. At the bottom, there are "← Back" and "Skip question →" links.
- Screen 3: Data**
Title: **Data**
Question: Do you need qualitative data, quantitative data or both?
Please select one:
 Qualitative data
 Quantitative data
 Both
 No preference
A large orange "Continue >" button is positioned below the options. At the bottom, there are "← Back" and "Skip question →" links.

Fig. 16: Decision process: more questions

The figure displays three sequential screenshots of the DEPRO decision process interface. Each screen features a header with the word "DEPRO" in green, a background image of a person working at a desk, and navigation options: "← Back" and "Skip question →".

Screen 1: Remotely
 Question: Do you need to work remotely?
 Please select one:
 Yes
 No
 No preference
 Continue >

Screen 2: Context
 Question: Do you want to observe users in their natural context?
 Please select one:
 Yes
 No
 No preference
 Continue >

Screen 3: Recap
 These are your answers to the previous questions. Please check them.
 Where are you in the development phase? Fully functional product
 What needs testing? Set of tasks
 How much time is available to you? Medium
 Do you need qualitative data, quantitative data or both? Both
 Do you need to work remotely? No preference
 Do you want to observe users in their natural context? No preference
 Show methods >

Fig. 17: Decision process: final questions and recap

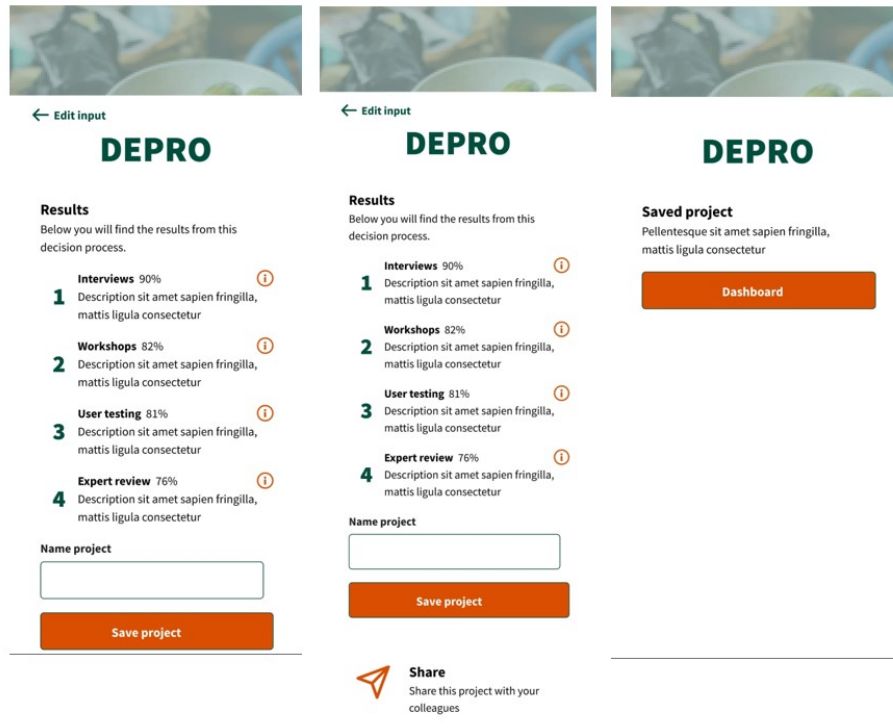


Fig. 18: Decision process: results of the decision process, share button and return to dashboard button

8 Testing and evaluation

In this thesis, a literature study combined with contextual inquiry and interviews have provided input for a UXEMs decision process tool. In the previous section, the decision tool is described. In this section, the evaluation of the prototype is discussed. This evaluation is summative for the purpose of this thesis, but can be used as a starting point for future work. This section describes process D: *testing and evaluation* as shown in figure 19.

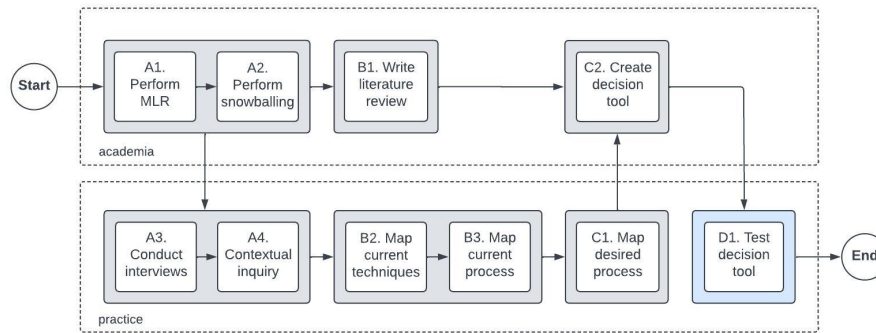


Fig. 19: Process D: testing and evaluation

8.1 Set up

The participants for this evaluation were team members of the UX team at location A. From this group, individuals were selected based on convenience sampling. In a short session, the users were asked to keep their most recent project in mind as a case for the decision tool. Four users of the UX team participated, with one team member being a junior employee, one being a medior employee and two being senior employees. All participants were familiar with Figma as a prototyping tool. All four participants were also observed or interviewed at some point during the contextual inquiry.

Participants were presented with DEPRO after a short introduction by the researcher about the subject and the present thesis. All participants were familiar with the thesis to some extent. After an introduction, the researcher starts the application and asks the participant to think aloud. As the participants had experience with the thinking aloud method, they did not find it difficult to narrate their thought process.

8.2 Results

All participants reported that they would use this tool if it were a fully functional product. At first most said they would not blindly trust DEPRO to make their decisions for

them, but they would use it to support and inform their own decisions. It does take slightly longer to go through the decision process than some participants would have hoped. Participants considered the user interface *simple but effective*. Especially for an internal application of which the user interface is not or rarely shown to external customers, the participants stated that UI does not matter much. Of course, the application should be easy to use and information should be easy to find. This did not seem to be a problem, as the application did not have a large number of features implemented.

"As long as it's usable and does what it should do, I think I would be content."

What participants were most interested in the future possibilities of a tool like this. The structured decision making can improve communication to customers, especially to apply methods that may be more expensive or time consuming but have a higher compatibility rating. The opportunities for improving contact between colleagues from different locations were also mentioned. One participant stated that they would find it interesting to know whether they possess the same knowledge at other locations. This sparked an idea for workshops or lunch talks to bring colleagues from multiple locations together and share knowledge.

As a functional prototype was tested, the quality of the results of the decision tool cannot be judged.

8.3 What now?

The theoretical foundation has been established and the prototype has been created. The evaluation proves that there are many opportunities for future work, but it depends on the company whether or not the tool will actually become a reality someday. Although the lack of uniformity persists, it is unclear whether the company can spare the resources to create a solution like the one proposed in this thesis.

9 Discussion

9.1 Methods

In this thesis, a multivocal literature review, contextual inquiry and interviews were used to find answers to the research questions. These sub research questions informed the creation of the the final deliverable. The contextual inquiry allowed for the collection of a broad set of data. It would have been challenging to collect the same amount of rich information with another method. The interviews provided more depth and knowledge about the existing decision methods, building on the data from the contextual inquiry.

9.2 Results

Table 8 shows that the expert evaluation, interviews, software tools and workshops were the methods that were most frequently known and used by the participants. The field studies and scales and questionnaires categories were used and known less. This could be due to field studies often taking a significant amount of time, which is typically not available in a business. Scales and questionnaires may have been used less as many scales and questionnaires have complicated, academic sounding names and were developed within academia, for academia. This data is not generalizable and results of contextual inquiry and interviews may differ depending on the company that is investigated.

9.3 Deliverable

If the tool would have been created as a web page, it would be less of an obstruction to the work flow. Most users of the tool would be in front of their PC anyway. This would only take away from the advantage that the tool can be used anywhere. It is also important to keep the information up-to-date and to keep an eye out for new methods or methods that may not be implemented into DEPRO yet, as the system does not update automatically. DEPRO cannot be generalized, as part of the research behind the tool is based on one company. Through careful consideration from the author, the influence of the company on this thesis is minimized. It is possible, however, that the DEPRO concept is applicable in other contexts after some modifications to the content.

The testing and evaluation showed a clear interest from the UX team members. From the interviews it had already become clear that some members of the UX teams would like a more structured approach to deciding which UX evaluation method to choose, and this was only confirmed during the testing and evaluation. While applying the thinking aloud evaluation method, the participants explored the prototype. The thinking aloud evaluation method was chosen as it is a low-cost method that can yield detailed results. It allows the user to explore the system on their own, without the help of the researcher. Another important factor in this decision was time constraints from the side of the participants, as the researcher could not take up too much of their time. The thinking aloud technique yielded interesting results. If more time had been available, perhaps a more detailed technique or exploration session would have been more suitable.

9.4 Implications

This thesis shines light on an otherwise vague concept: the UXEM decision process. The DEPRO tool is introduced to structure this decision process. This problem does not just exist in practice, but also in the literature. The literature review shows that very little similar systems exist. A tool that takes the user through the process in a structured way adds explainability and reproducibility to research. It shows that the researchers have considered other options and the chosen methods are, to their knowledge, the best suitable for their goal. For academic purposes, it may not always be necessary to enter the known and used methods.

9.5 Future work

The future of this decision support tool is unclear. The non-uniformity of knowledge within the company persists, but creating the proposed tool means temporarily sacrificing much needed resources. The testing and evaluation yielded interesting insights that could prove useful for future research. Additionally, there are still some requirements that have not been implemented. In the future, the application could be made into a fully functional product. A functionality to recommend triangulation could be added. The tool could be expanded to cover all user experience methods, not just evaluation methods. To improve generalizability, a broader study of the industry should be conducted. To account for certain (academic) purposes, a version without the used and known methods could be created. DEPRO could, in the future, serve as the bridge between academic UX and industry UX.

10 Conclusion

Decentralized cognition can be a problem in large companies, but especially in companies who have recently gone through organizational changes. This problem exists in the company introduced in this paper where several smaller companies were recently merged under one large name. In this thesis, DEPRO is proposed as a tool to improve centralized cognition by structuring information in a uniform way.

In the first section of this thesis, the following sub research questions were introduced:

- SQ1: What types of user experience evaluation methods currently exist in both academic and corporate settings?
- SQ2: What are the positive and negative aspects of each type of user experience evaluation method?
- SQ3: Which of the found user experience evaluation methods are best suited for academic purposes and which are best suited for corporate purposes?
- SQ4: What decision processes for user experience evaluation methods currently exist in academic and corporate settings?
- SQ5: What are the requirements for an interactive decision support system intended to aid the decision process for user experience evaluation methods?

In the literature study, SQ1 was answered through a long list of UXEMs that were found in the literature (Appendix A). Additionally, it was found that many classifications of UX exist, all focusing on different aspects such as the user, the time required, qualitative or quantitative and more. For the purpose of this thesis, seven categories of UX were identified through examining trends in the MLR table: expert evaluations, field studies, interviews, measurements, scales and questionnaires, software tools and workshops. SQ2 could also be answered through the literature review. Each aforementioned category has a unique set of challenges and benefits, but the overarching categories had some aspects in common. Examples of commonly found positive and negative aspects were low-cost methods, methods requiring an experienced researcher or data analysis being a long and difficult process. With the literature review, SQ3 could be answered as well. Methods that are preferred in the industry are often low-cost, informal methods that can be applied to working prototypes. In the case of academic methods being adopted into the industry, some changes may need to be made. Methods that yield quantitative results are, according to the literature study, the best suited to apply in industry settings. Some tools and models exist that are somewhat similar to the tool proposed in this thesis, but a product that is truly similar in all aspect of the tool could not be found. To answer SQ4, some decision processes exist, but they are very limited or inaccessible. Through contextual inquiry and interviews, eight user stories were constructed. Based on these user stories, requirements were created and SQ5 was answered. These requirements were analysed in a MoSCoW analysis and arranged based on importance.

The answers to these questions provide an answer to the main research question:

- How can challenges that come with choosing user experience evaluation methods in a non-uniform way in a large end-to-end agency be solved through an interactive decision model?

The challenges that come with choosing UXEMs in a non-uniform way can in the case of this thesis be mitigated through a tool that structures the decision process: DEPRO. With the use of the tool, all users can find the most optimal UXEM for them, based on personal skills and external factors such as time constraints. DEPRO can increase reproducibility and explainability of results in both academic and industry settings.

References

1. Gotik-methode: wat is het en wat zijn de voordelen? (2019), <https://www.gww-bouw.nl/artikel/gotik-methode-wat-is-het-en-wat-zijn-de-voordelen/>
2. Abeele, V.V., Zaman, B.: Laddering the user experience. In: User experience evaluation methods in product development (UXEM'09)-workshop. pp. 1–5. Citeseer (2009)
3. Abuaddous, H.Y., Saleh, A.M., Enaizan, O., Ghabban, F., Al-Badareen, A.B.: Automated user experience (ux) testing for mobile application: Strengths and limitations. *International Journal of Interactive Mobile Technologies* **16**(4) (2022)
4. Albert, B., Tullis, T.: *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes (2013)
5. Alhadreti, O., Mayhew, P.: Rethinking thinking aloud: A comparison of three think-aloud protocols. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–12 (2018)
6. Alroobaea, R., Mayhew, P.J.: How many participants are really enough for usability studies? In: 2014 Science and Information Conference. pp. 48–56. IEEE (2014)
7. Alves, R., Valente, P., Nunes, N.J.: The state of user experience evaluation practice. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational. pp. 93–102 (2014)
8. Ardito, C., Buono, P., Caivano, D., Costabile, M.F., Lanzilotti, R.: Investigating and promoting ux practice in industry: An experimental study. *International Journal of Human-Computer Studies* **72**(6), 542–551 (2014)
9. Auger, C.P.: *Information Sources in Grey Literature*. Bowker-Saur, 4th edn. (1998)
10. Baig, M.Z., Kavakli, M.: A survey on psycho-physiological analysis & measurement methods in multimodal systems. *Multimodal Technologies and Interaction* **3**(2), 37 (2019)
11. Bakker, S., Markopoulos, P., De Kort, Y.: Opos: an observation scheme for evaluating head-up play. In: Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges. pp. 33–42 (2008)
12. Bang, K., Kanstrup, M.A., Kjems, A., Stage, J.: Adoption of ux evaluation in practice: An action research study in a software organization. In: *Human-Computer Interaction—INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25–29, 2017, Proceedings, Part IV* 16. pp. 169–188. Springer (2017)
13. Bernhaupt, R.: User experience evaluation methods in the games development life cycle. In: *Game user experience evaluation*, pp. 1–8. Springer (2015)
14. Bevan, N.: Classifying and selecting ux and usability measures. In: *International Workshop on Meaningful Measures: Valid Useful User Experience Measurement*. vol. 11, pp. 13–18. Institute of Research in Informatics of Toulouse (IRIT) Toulouse, France (2008)
15. Bevan, N., Macleod, M.: Usability measurement in context. *Behaviour & information technology* **13**(1-2), 132–145 (1994)
16. Beyer, H., Holtzblatt, K.: Contextual design. *interactions* **6**(1), 32–42 (1999)
17. Boren, T., Ramey, J.: Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication* **43**(3), 261–278 (2000)
18. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* **25**(1), 49–59 (1994)
19. Buche, A., Chandak, D., Zadgaonkar, A.: Opinion mining and analysis: a survey. arXiv preprint arXiv:1307.3336 (2013)
20. Cavalcante¹, E., Rivero¹, L., Conte¹, T.: Max: A method for evaluating the post-use user experience through cards and a board. In: *27th International Conference on Software Engineering and Knowledge Engineering (SEKE 2015)*. pp. 495–500 (2015)

21. Chang, Y.n., Lim, Y.k., Stolterman, E.: Personas: from theory to practices. In: Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges. pp. 439–442 (2008)
22. Charoenpruksachat, A., Longani, P.: Comparative study of usability evaluation methods on a hyper casual game. In: 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering. pp. 153–156. IEEE (2021)
23. Chen, H., Zimbra, D.: Ai and opinion mining. *IEEE Intelligent Systems* **25**(3), 74–80 (2010)
24. Cheng, F., Yu, S., Qin, S., Chu, J., Chen, J.: User experience evaluation method based on online product reviews. *Journal of Intelligent & Fuzzy Systems* **41**(1), 1791–1805 (2021)
25. Chung, T.K., Sahari, N.: Utilitarian or experiential? an analysis of usability questionnaires. *International Journal of Computer Theory and Engineering* **7**(2), 167–171 (2015)
26. Cirkovic, S.: Grey literature—the chameleon of information resources. *Infototeca-Journal For Digital Humanities* **18**(1), 75–83 (2018)
27. Darin, T., Coelho, B., Borges, B.: Which instrument should i use? supporting decision-making about the evaluation of user experience. In: International conference on human-computer interaction. pp. 49–67. Springer (2019)
28. Davies, M.: Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? *Higher education* **62**, 279–301 (2011)
29. De Matos, P., Cham, J.A., Cao, H., Alcántara, R., Rowland, F., Lopez, R., Steinbeck, C.: The enzyme portal: a case study in applying user-centred design methods in bioinformatics. *BMC bioinformatics* **14**, 1–15 (2013)
30. De Souza, C.S., Leitão, C.F., Prates, R.O., Da Silva, E.J.: The semiotic inspection method. In: Proceedings of VII Brazilian symposium on Human factors in computing systems. pp. 148–157 (2006)
31. Dell’Era, C., Landoni, P.: Living lab: A methodology between user-centred design and participatory design. *Creativity and Innovation Management* **23**(2), 137–154 (2014)
32. Den Uyl, M., Van Kuilenburg, H.: The facereader: Online facial expression recognition. In: Proceedings of measuring behavior. vol. 30, pp. 589–590. Citeseer (2005)
33. Desmet, P.: Measuring emotion: Development and application of an instrument to measure emotional responses to products. *Funology: From usability to enjoyment* pp. 111–123 (2005)
34. Dhoub, A., Assila, A., Trabelsi, A., Kolski, C., Neji, M.: Factors affecting the choice of usability evaluation methods for interactive adaptive systems. In: International Conference on Human-Centred Software Engineering. pp. 270–282. Springer (2018)
35. Dhoub, A., Trabelsi, A., Kolski, C., Neji, M.: A classification and comparison of usability evaluation methods for interactive adaptive systems. In: 2016 9th International Conference on Human System Interactions (HSI). pp. 246–251. IEEE (2016)
36. Falkowska, J., Sobecki, J., Pietrzak, M.: Eye tracking usability testing enhanced with eeg analysis. In: Design, User Experience, and Usability: Design Thinking and Methods: 5th International Conference, DUXU 2016, Held as Part of HCI International 2016, Toronto, Canada, July 17–22, 2016, Proceedings, Part I 5. pp. 399–411. Springer (2016)
37. Fernandez, A., Abrahão, S., Insfran, E.: A systematic review on the effectiveness of web usability evaluation methods (2012)
38. Fernandez, A., Insfran, E., Abrahão, S.: Usability evaluation methods for the web: A systematic mapping study. *Information and software Technology* **53**(8), 789–817 (2011)
39. Ferreira, B.M., Rivero, L., Valentim, N.M.C., Zilse, R., Koster, A., Conte, T.: Evaluation of ux methods: Lessons learned when evaluating a multi-user mobile application. In: International Conference on Human-Computer Interaction. pp. 279–290. Springer (2016)

40. Filippi, S.: Persel, a ready-to-use personality-based user selection tool to maximize user experience redesign effectiveness. *Multimodal Technologies and Interaction* **4**(2), 13 (2020)
41. Finstad, K.: The usability metric for user experience. *Interacting with computers* **22**(5), 323–327 (2010)
42. Fischer, H., Kauer-Franz, M., Winter, D., Latt, S.: Uux method selection. *i-com* **15**(1), 111–116 (2016)
43. Fischer, H., Streng, B., Nebe, K.: Towards a holistic tool for the selection and validation of usability method sets supporting human-centered design. In: *International Conference of Design, User Experience, and Usability*. pp. 252–261. Springer (2013)
44. Garousi, V., Felderer, M., Mäntylä, M.V.: Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* **106**, 101–121 (2019). <https://doi.org/https://doi.org/10.1016/j.infsof.2018.09.006>, <https://www.sciencedirect.com/science/article/pii/S0950584918301939>
45. Gaver, B., Dunne, T., Pacenti, E.: Design: cultural probes. *interactions* **6**(1), 21–29 (1999)
46. Georges, V., Courtemanche, F., Sénécal, S., Léger, P.M., Nacke, L., Pourchon, R.: The adoption of physiological measures as an evaluation tool in ux. In: *HCI in Business, Government and Organizations. Interacting with Information Systems: 4th International Conference, HCIBGO 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 4*. pp. 90–98. Springer (2017)
47. Gerken, J., Jetter, H.C., Zöllner, M., Mader, M., Reiterer, H.: The concept maps method as a tool to evaluate the usability of apis. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. pp. 3373–3382 (2011)
48. Goffin, K., Lemke, F., Koners, U., Goffin, K., Lemke, F., Koners, U.: Repertory grid technique. *Identifying Hidden Needs: Creating Breakthrough Products* pp. 125–152 (2010)
49. Gordillo, A., Barra, E., Aguirre, S., Quemada, J.: The usefulness of usability and user experience evaluation methods on an e-learning platform development from a developer's perspective: A case study. In: *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. pp. 1–8. IEEE (2014)
50. Grigoreanu, V., Mohanna, M.: Informal cognitive walkthroughs (icw) paring down and pairing up for an agile world. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 3093–3096 (2013)
51. Hartson, R., Pyla, P.S.: *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier (2012)
52. Hasan, L., Morris, A., Probst, S.: A comparison of usability evaluation methods for evaluating e-commerce websites. *Behaviour & Information Technology* **31**(7), 707–737 (2012)
53. Hassenzahl, M., Wessler, R.: Capturing design space from a user perspective: The repertory grid technique revisited. *International Journal of Human-Computer Interaction* **12**(3-4), 441–459 (2000)
54. Hektner, J.M., Schmidt, J.A., Csikszentmihalyi, M.: *Experience sampling method: Measuring the quality of everyday life*. Sage (2007)
55. Hinderks, A., Schrepp, M., Mayo, F.J.D., Escalona, M.J., Thomaschewski, J.: Developing a ux kpi based on the user experience questionnaire. *Computer Standards & Interfaces* **65**, 38–44 (2019)
56. Holzinger, A.: Usability engineering methods for software developers. *Communications of the ACM* **48**(1), 71–74 (2005)
57. Hussain, J., Ul Hassan, A., Muhammad Bilal, H.S., Ali, R., Afzal, M., Hussain, S., Bang, J., Banos, O., Lee, S.: Model-based adaptive user interface based on context and user experience evaluation. *Journal on Multimodal User Interfaces* **12**(1), 1–16 (2018)
58. Hwang, W., Salvendy, G.: Number of people required for usability evaluation: the 10±2 rule. *Communications of the ACM* **53**(5), 130–133 (2010)

59. Inan Nur, A., B. Santoso, H., O. Hadi Putra, P.: The method and metric of user experience evaluation: A systematic literature review. In: 2021 10th International Conference on Software and Computer Applications. pp. 307–317 (2021)
60. Ishikawa, D., Kato, T., Kita, C.: A comparative analysis of usability evaluation methods on their versatility in the face of diversified user input methods. In: HCI International 2015- Posters' Extended Abstracts: International Conference, HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015. Proceedings, Part I. pp. 32–37. Springer (2015)
61. ISO: ISO9241-210:2019 (2019), <https://www.iso.org/standard/77520.html>
62. John, B.E., Kieras, D.E.: Using goms for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction (TOCHI)* **3**(4), 287–319 (1996)
63. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.B.: User experience over time: an initial framework. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 729–738 (2009)
64. Kato, T.: What “question-asking protocols” can say about the user interface. *International Journal of Man-Machine Studies* **25**(6), 659–673 (1986)
65. Kelly, G.A.: The psychology of personal constructs. Volume 1: A theory of personality. WW Norton and Company (1955)
66. Kieras, D.: A guide to goms model usability evaluation using ngomsl. In: Handbook of human-computer interaction, pp. 733–766. Elsevier (1997)
67. Kim, J.H., Gunn, D.V., Schuh, E., Phillips, B., Pagulayan, R.J., Wixon, D.: Tracking real-time user experience (true) a comprehensive instrumentation solution for complex systems. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. pp. 443–452 (2008)
68. Kirakowski, J.: The use of questionnaire methods for usability assessment (1994)
69. Kirakowski, J., Corbett, M.: Sumi: the software usability measurement inventory. *British Journal of Educational Technology* **24**(3), 210–212 (1993)
70. Korhonen, H., Arrasvuori, J., Väänänen-Vainio-Mattila, K.: Let users tell the story: evaluating user experience with experience reports. In: CHI'10 Extended Abstracts on Human Factors in Computing Systems, pp. 4051–4056 (2010)
71. Kort, J., Vermeeren, A., Fokker, J.E.: Conceptualizing and measuring user experience. Towards a UX manifesto p. 57 (2007)
72. Krippendorff, K., Butter, R.: Product semantics-exploring the symbolic qualities of form. *Departmental Papers (ASC)* p. 40 (1984)
73. Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., Sinnelä, A.: Ux curve: A method for evaluating long-term user experience. *Interacting with computers* **23**(5), 473–483 (2011)
74. Kujala, S., Walsh, T., Nurkka, P., Crisan, M.: Sentence completion for understanding users and evaluating user experience. *Interacting with Computers* **26**(3), 238–255 (2014)
75. Kurosu, M., Hashizume, A., Ueno, Y.: User experience evaluation by erm: experience recollection method. In: International Conference on Human-Computer Interaction. pp. 138–147. Springer (2018)
76. Lachner, F., Naegelein, P., Kowalski, R., Spann, M., Butz, A.: Quantified ux: Towards a common organizational understanding of user experience. In: Proceedings of the 9th Nordic conference on human-computer interaction. pp. 1–10 (2016)
77. Lavie, T., Tractinsky, N.: Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies* **60**(3), 269–298 (2004)
78. Li, M., Albayrak, A., Zhang, Y., Eijk, D.v., Yang, Z.: Comparison of questionnaire based and user model based usability evaluation methods. In: Congress of the International Ergonomics Association. pp. 1081–1098. Springer (2018)

79. Liapis, A., Katsanos, C., Karousos, N., Xenos, M., Orphanoudakis, T.: User experience evaluation: A validation study of a tool-based approach for automatic stress detection using physiological signals. *International Journal of Human-Computer Interaction* **37**(5), 470–483 (2021)
80. Liikkanen, L.A., Reavey, H.: Resonance testing: an industry approach for experiential concept evaluation. *International Journal of Product Development* **20**(4), 265–285 (2015)
81. Macleod, M., Bowden, R., Bevan, N., Curson, I.: The music performance measurement method. *Behaviour & Information Technology* **16**(4-5), 279–293 (1997)
82. Madsen, M., Gregor, S.: Measuring human-computer trust. In: 11th australasian conference on information systems. vol. 53, pp. 6–8. Citeseer (2000)
83. Magües, D.A., Fonseca C, E.R., Castro, J.W., Acuña, S.T.: Usability evaluation methods adopted in agile development processes. p. 423 – 436 (2018), <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054101645&partnerID=40&md5=08068cb47abfd152698ad88e2248cff9>
84. Maia, C.L.B., Furtado, E.S.: Retuxe: A framework for user’s emotional evaluation based on psychophysiological measures. In: Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems. pp. 1–4 (2017)
85. Maia, C.L.B., Furtado, E.S.: A systematic review about user experience evaluation. In: International conference of design, user experience, and usability. pp. 445–455. Springer (2016)
86. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology* **25**(2), 141–158 (2006)
87. Mattelmäki, T., et al.: Design probes. Aalto University (2006)
88. Melo, P., Jorge, L.: Quantitative support for ux methods identification: how can multiple criteria decision making help? *Universal Access in the Information Society* **14**(2), 215–229 (2015)
89. Mugge, R., Schifferstein, H.N., Schoormans, J.P.: A longitudinal study of product attachment and its determinants. *ACR European Advances* (2005)
90. Najmiec, A., Zawieska, W.M., Suchecka, M., Kurowski, J.: Ergonomic aspects of using a computer system for hazard registration and occupational risk assessment. *Systems, Social, and Internationalization Design Aspects of Human-computer Interaction: Volume 2*, 458 (2001)
91. Nakamura, W.T., Ahmed, I., Redmiles, D., Oliveira, E., Fernandes, D., de Oliveira, E.H., Conte, T.: Are ux evaluation methods providing the same big picture? *Sensors* **21**(10), 3480 (2021)
92. Nielsen, J.: Ten usability heuristics (2005)
93. Oyugi, C., Abdelnour-Nocera, J., Clemmensen, T.: Harambee: a novel usability evaluation method for low-end users in kenya. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational. pp. 179–188 (2014)
94. Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., McKenzie, J.E.: Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* **372** (2021). <https://doi.org/10.1136/bmj.n160>, <https://www.bmj.com/content/372/bmj.n160>
95. Paz, F., Paz, F.A., Pow-Sang, J.A.: Application of the communicability evaluation method to evaluate the user interface design: a case study in web domain. In: Design, User Experience, and Usability: Design Thinking and Methods: 5th International Conference, DUXU 2016,

- Held as Part of HCI International 2016, Toronto, Canada, July 17–22, 2016, Proceedings, Part I 5. pp. 479–490. Springer (2016)
96. Paz, F., Pow-Sang, J.A.: Current trends in usability evaluation methods: a systematic review. In: 2014 7th International Conference on Advanced Software Engineering and Its Applications. pp. 11–15. IEEE (2014)
 97. Paz, F., Pow-Sang, J.A.: Usability evaluation methods for software development: a systematic mapping review. In: 2015 8th International Conference on Advanced Software Engineering & Its Applications (ASEA). pp. 1–4. IEEE (2015)
 98. Paz, F., Pow-Sang, J.A.: A systematic mapping review of usability evaluation methods for software development process. *International Journal of Software Engineering and Its Applications* **10**(1), 165–178 (2016)
 99. Pernice, K.: User interviews: How, when, and why to conduct them (Oct 2018), <https://www.nngroup.com/articles/user-interviews/>
 100. Pettersson, I., Lachner, F., Frison, A.K., Riener, A., Butz, A.: A bermuda triangle? a review of method application and triangulation in user experience evaluation. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–16 (2018)
 101. Prates, R.O., De Souza, C.S., Barbosa, S.D.: Methods and tools: a method for evaluating the communicability of user interfaces. *interactions* **7**(1), 31–38 (2000)
 102. Quiñones, D., Rusu, C., Rusu, V.: A methodology to develop usability/user experience heuristics. *Computer standards & interfaces* **59**, 109–129 (2018)
 103. Rajeshkumar, S., Omar, R., Mahmud, M.: Taxonomies of user experience (ux) evaluation methods. In: 2013 International Conference on Research and Innovation in Information Systems (ICRIIS). pp. 533–538. IEEE (2013)
 104. Resnick, M., Elkerton, J., Maher, P., Pastel, R., Rodriguez, A., Kelley, J.: Triangulation of multiple human factors methods in user experience design and evaluation. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 57, pp. 404–408. SAGE Publications Sage CA: Los Angeles, CA (2013)
 105. Ribeiro, T., de Souza, P.: A study on the use of personas as a usability evaluation method. In: Proceedings of the 16th International Conference on Enterprise Information Systems-Volume 3. pp. 168–175 (2014)
 106. Rico-Olarte, C., López, D.M., Kepplinger, S.: Towards a conceptual framework for the objective evaluation of user experience. In: International Conference of Design, User Experience, and Usability. pp. 546–559. Springer (2018)
 107. Rivero, L., Conte, T.: Using a study to assess user experience evaluation methods from the point of view of users. vol. 3, p. 88 – 95 (2015). <https://doi.org/10.5220/0005377300880095>, <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84939547869&doi=10.5220%2f0005377300880095&partnerID=40&md5=ced095329949b6fde0543d84f6f646c7>, cited by: 1; All Open Access, Hybrid Gold Open Access
 108. Rohrer, C.: When to use which user-experience research methods. Nielsen Norman Group **12**, 21 (2014)
 109. Salazar, K.: Contextual inquiry: Inspire design by observing and interviewing users in their context (Dec 2020), <https://www.nngroup.com/articles/contextual-inquiry/>
 110. Schifferstein, H.N., Zwartkruis-Pelgrim, E.P.: Consumer-product attachment: Measurement and design implications. *International journal of design* **2**(3) (2008)
 111. Schopf, J.: Towards a prague definition of grey literature. In: Twelfth International Conference on Grey Literature : GL12 Conference Proceedings : Transparency in Grey Literature (2010)

112. Schrepp, M., Thomaschewski, J.: Design and validation of a framework for the creation of user experience questionnaires (2019)
113. Schütte, S.: Towards a common approach in kansei engineering. a proposed model. (2007)
114. Scollon, C.N., Kim-Prieto, C., Diener, E.: Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness studies* **4**(1), 5–34 (2003)
115. Seyff, N., Ollmann, G., Bortenschlager, M.: Appecho: a user-driven, in situ feedback approach for mobile platforms and applications. In: *Proceedings of the 1st International Conference on Mobile Software Engineering and Systems*. pp. 99–108 (2014)
116. Spool, J., Schroeder, W.: Testing web sites: Five users is nowhere near enough. In: *CHI'01 extended abstracts on Human factors in computing systems*. pp. 285–286 (2001)
117. Ten, A.C., Paz, F.: A systematic review of user experience evaluation methods in information driven websites. In: *International Conference of Design, User Experience, and Usability*. pp. 492–506. Springer (2017)
118. Thompson, E.R.: Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology* **38**(2), 227–242 (2007)
119. Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., Bergel, M.: An empirical comparison of lab and remote usability testing of web sites. In: *Usability Professionals Association Conference* (2002)
120. Väänänen-Vainio-Mattila, K., Roto, V., Hassenzahl, M.: Towards practical user experience evaluation methods. *Meaningful measures: Valid useful user experience measurement (VUUM)* pp. 19–22 (2008)
121. Väänänen-Vainio-Mattila, K., Roto, V., Hassenzahl, M.: Towards practical user experience evaluation methods. *Meaningful measures: Valid useful user experience measurement (VUUM)* pp. 19–22 (2008)
122. Vaananen-Vainio-Mattila, K., Waljas, M.: Evaluating user experience of cross-platform web services with a heuristic evaluation method. *International Journal of Arts and Technology* **3**(4), 402–421 (2010)
123. Valencia, K., Botella, F., Rusu, C.: A property checklist to evaluate the user experience for people with autism spectrum disorder. In: *Social Computing and Social Media: Design, User Experience and Impact: 14th International Conference, SCSM 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I*. pp. 205–216. Springer (2022)
124. Van Den Haak, M., De Jong, M., Jan Schellens, P.: Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology* **22**(5), 339–351 (2003)
125. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: Toward a unified view. *MIS quarterly* pp. 425–478 (2003)
126. Vermeeren, A., Kort, J., Cremers, A., Fokker, J.: Comparing ux measurements, a case study. In: *Proceedings of the International Workshop on Meaningful Measures: Valid Useful Experience Measurement*, Reykjavik, Iceland, June. vol. 18, pp. 72–78 (2008)
127. Vermeeren, A.P., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K.: User experience evaluation methods: current state and development needs. In: *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*. pp. 521–530 (2010)
128. Virzi, R.A.: Refining the test phase of usability evaluation: How many subjects is enough? *Human factors* **34**(4), 457–468 (1992)
129. Wang, J., Wang, X., Lu, J., Xu, Z.: Investigating the user experience of mind map software: A comparative study based on eye tracking. *International Journal of Advanced Computer Science and Applications* **13**(11) (2022)

130. Weichbroth, P.: Usability of mobile applications: a systematic literature study. *Ieee Access* **8**, 55563–55577 (2020)
131. Wilson, C.: Interview techniques for UX practitioners: A user-centered design method. Newnes (2013)
132. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: A presence questionnaire. *Presence* **7**(3), 225–240 (1998)
133. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering. pp. 1–10 (2014)
134. Yong, L.T.: User experience evaluation methods for mobile devices. In: Third International Conference on Innovative Computing Technology (INTECH 2013). pp. 281–286. IEEE (2013)
135. Zaina, L.A., Sharp, H., Barroca, L.: Ux information in the daily work of an agile team: A distributed cognition analysis. *International Journal of Human-Computer Studies* **147**, 102574 (2021)
136. Zaman, B., Abeele, V.V.: Laddering with young children in user experience evaluations: theoretical groundings and a practical case. In: Proceedings of the 9th International Conference on Interaction Design and Children. pp. 156–165 (2010)
137. Zarour, M., Alharbi, M.: User experience framework that combines aspects, dimensions, and measurement methods. *Cogent Engineering* **4**(1), 1421006 (2017)
138. Zhang, Z., Basili, V., Shneiderman, B.: Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Software Engineering* **4**, 43–69 (1999)

A MLR UX evaluation methods

Table 12: Expert evaluation methods from the MLR

Method	Found in	Description
Cognitive jogthrough	[98]	Alternative version of cognitive walkthrough where the observer asks themselves a set of questions from the perspective of the user. The answers to these questions are ranked according to the number of users who are expected to struggle with this [98]. More room for discussions between members of the evaluation team /citerowley1992cogjog.
Cognitive task analysis	[38][98] [130]	Researcher observes an ordinary user's interaction with a system to form an understanding of how tasks are performed and goals are achieved. Can help with identifying tasks a system must have and focuses on understanding decision-making, problem-solving, memory, attention and judgement [98].
Cognitive walkthrough	[35][37] [38][58] [96][97] [98][130]	Usability expert simulates the actions of a new user while the inspector identifies potential issues of usability [98].
Domain specific inspection	[98]	DSI was designed to assess and improve usability of social network websites, but could be applied to other domains as well. Depending on the domain the researchers need, relevant areas and attributes should be determined.
Expert review	[37][117] [137][127]	Experts review a design based on heuristics, guidelines and their own expert knowledge. The experts identify each usability problem, rate it based on severity and recommend a solution for the problem [127]. Can be combined with usability testing for more thorough results.
Goals, operators, methods and selection (GOMS) rules analysis	[38][130]	Human task performance is modeled in terms of Goals, Operators, Methods and Selection rules to predict execution and learning time [38]. The general GOMS concept is described as "it is useful to analyze the knowledge of how to do a task in terms of goals, operators, methods and selection rules" [62]. There are many different GOMS models, each suitable for a different purpose [66].

Heuristic evaluation & guideline review	[13][35] [37][38] [58][83] [96][97] [98][122] [130]	When using heuristic evaluation, usability experts judge a software system based on established usability principles (heuristics) [?].
Participatory heuristic evaluation	[98]	Participatory heuristic evaluation uses the largely same principles as traditional heuristic evaluation, but it involves the participation of end users as domain expert inspectors [98].
Personas	[96][98] [105][137]	Personas are the descriptions of fictitious users of the system, where their characteristics and goals are emphasized [21,98]. In the context of UXEM, evaluating using personas would include analyzing the user interface while considering the goals, attitudes, behaviours and (business) objectives of the fictitious user [98].
Playability heuristics	[127]	Using playability heuristics, the playability aspect of a game can be evaluated [127]. May be applied in an expert evaluation setting where the experts should familiarize themselves with the heuristics beforehand. This method is quick and cheap and can be applied to many stages in the development process of the game (early or late).
Property checklists	[49][127]	A structured form of expert evaluation where the expert reviews a checklist consisting of design goals for different properties of a product (e.g. colour, sounds, materials, form, graphics, interaction design, functionality) [127]. The goal of this inspection method [38] is to verify the presence or absence of properties to ensure everything has been considered or completed [123]. This method is quick and does not require any participants aside from the expert. However, the expert is not the user.
Task environment analysis	[38][130]	Evaluation or assessment of the mapping between users' goals and user interface tasks [38]. This method is typically classified under analytical modelling [25,38].

Usability guidelines	[98][117]	A group of specialists evaluate a graphical interface of a product according to pre-defined usability guidelines [98]. This method is similar to heuristic evaluation, but the procedure is different. Here, each specialist can work individually. The guidelines do not need to be rated on the severity and criticality of each issue. This method does not necessarily require a set of usability heuristics, even guidelines provided by the software development company can suffice.
User workflow	[98]	Diagrams represent the available paths in a system to complete a certain task [98]. Workflows can also be used for hierarchical task analysis [29] and different user's preferences can be identified. This method can be used for processes that may include many sub-tasks.
Web usability evaluation process	[98]	Paz et al. [98] developed this protocol because they found that there was no formal procedure that allowed specialists to evaluate the usability of software products through heuristic inspection in a structured way. They found that Nielsen's traditional heuristics [92] did not cover new aspects such as real-time processing and sophisticated designs. This method is designed to fill those gaps.

Table 13: Field study methods from the MLR

Method	Found in	Description
Day reconstruction method (DRM)	[127]	During a field study, DRM can be applied as a self-reported measure. At the end of the day, the participants are asked to report all activities related to the use of the product, recording a name and an estimate of time spent. Then they move on to the second part: experience narration. Here, the participants are asked to pick the three most impactful experiences, either satisfying or dissatisfying. For each experience, the participants are asked to write a story to describe the situation in detail, their feelings and their perceptions of the product in the moment [127,63].
Diary study	[127][137]	Participants are asked to report data about the use of a product over a longer period of time.

Experience clip	[127]	Two users who know each other well are paired up in their natural environment. They are asked to interact with a mobile application and to shoot clips of this interaction. Participants are encouraged to shoot as many clips as possible, to engage in a discussion and to elaborate on their experiences [127].
Experience report	[137]	Experience reports are open-ended experience stories written by users after using their products in real contexts of use [70].
Experience sampling method (ESM)	[127][134]	With this method, the researcher can collect experiences as they happen. This eliminates the risk of memory effects. Participants receive a specific device to answer ESM questions on [?].
Experiential contextual inquiry	[127]	The researcher takes the role of an apprentice as they observe the user in real context. They ask questions related to their use of the system. The researcher pays special attention to the emotional aspect of use (i.e. what elicits a positive and negative emotion) [127].
Immersion	[127]	The researcher immerses themselves in the system by using the system in real life contexts. They evaluate their use of the system. With this method, the researcher is the participant in their own field study [127]. This method is a quick and lightweight method for finding experience, technical and usability bugs and can be used for early or functional prototypes.
Living lab method	[127]	Living labs allow the testing of complex solutions in multiple real life contexts [127]. When using this method, researchers can study the users' behaviour in naturalistic living environments, which can help better understand how to create technologies that fit the complexity of (everyday) life. This method is applied over time and is time and resource intensive and should be combined with other methods in the field. The living lab method has two primary elements: it involves a real-life test and experimentation environment and users who know they are co-involved in the innovation process [31].

Long term diary study	[127]	A long term diary study can capture the user experience of a product over a longer period of time, e.g. six months to a year [127]. The participants are given a prototype of a product and report their experiences and emotions at fixed intervals. This is done in the form of journal entries. This method can shine light on the user experience in different settings, but can be perceived as intensive and thus may have a high drop-out rate. To apply this method, working prototypes are required.
Timed ESM	[127]	Participants can either report their experience at a predefined point in time, or are prompted by the system [127]. The user is then asked to report different kinds of data, e.g. what feelings were triggered in the previous interactions, what they feel right now or how they feel about the system overall. The format for this data can range from questionnaire choices to audio recording, image or video. This method is a type of field research that allows researchers to collect experience data without being with the participant. The method collects retrospective information. When the user is prompted to report their experience, they might not be present. The trigger might also interrupt the user in their flow and trigger negative emotions [54].

Table 14: Interview methods from the MLR

Method	Found in	Description
Audio narrative	[127]	Users narrate their experiences with the product, which is recorded through audio. Free story format.
Contextual laddering	[127]	Interviewing technique for qualitative data gathering and quantitative data analysis technique, preferably done in context. The aim is to find out the reasons why attributes are liked or important to reveal the participant's dominant attributes - consequences - values chains related to a product [127].
Exploration test	[127]	This is an ethnographic test to evaluate a user's perception of a design. The researcher shows the user a design or prototype to gain their perception, and asks about other, similar products they use or other ways they complete the task.

Interview	[13][22] [37][38] [49][85] [96][97] [98][130] [137]	In an interview, the end user and usability expert engage in a discussion about the usability of a software system [98]. This interview may be structured, semi-structured or unstructured.
Private camera conversation	[127]	This method exists to mitigate interviewer bias [127]. The participant reports to a camera in private, either during the use of a product or afterwards. This method might yield different responses than a typical face-to-face interview, as the data might be more authentic. When using the private camera conversation method, there is no guarantee that the interviewee talks about the desired topics. Participants may also feel uncomfortable talking to a camera.
Product semantic analysis (PSA)	[127]	Product semantic analysis has a basis in product semantics, which concerns the relationship between the user and product on one side and the importance that artefacts assume in a social and operational context on the other side [72,127]. Assesses the visual aspect of a product's properties (e.g. lightness or softness). When using this method, relevant words need to be identified through interviews first. After these interviews, a semantic scale is constructed which includes preferred and non-preferred words. The scale has two points with a neutral middle (max. value - neutral - min. value). Only one of the adjectives is used, the other side is referred to as 'the opposite'. After the scale has been constructed, more interviews take place with consumers to create a desired product semantic profile to show the desired strengths of expressions. In the end, the consumers have rated the perceived visual aspect of a product using this method [127].
Semi-structured experience interview	[127]	In a semi-structured experience interview, predefined questions are combined with open-ended exploration. This type of interview is done when there is some knowledge regarding the investigated topic, but additional details are required [131]. Preparation from the researcher is required. The interviews provide rich data, and when using this type of evaluation method, a small sample size is sufficient. The analysis can be time consuming [127].

Table 15: Measurement methods from the MLR

Method	Found in	Description
Physiological signals / physiological UX evaluation / psychophysiological measurements / physiological arousal via electrodermal activity / facial EMG	[13][59] [84][106] [127][134]	Physiological signals (e.g. facial muscles, skin perspiration, heart beat) are the convergence point between the physical state of the user and the measurement of emotions, and thus is an objective UX evaluation method [106]. Physiological reactions of a participant are recorded using sensors such as skin conductance (SC), electroencephalography (EEG), electrocardiography (ECG) [10] or electromyography (EMG) [127]. This objective method may be combined with a self-reported method to find out what the user has experienced [46,86,127]. Tools for combining and analyzing measurement data exist (e.g. PhysiOBS [79]). The measurement method may limit movement for the participant.

Table 16: Scale and questionnaire methods from the MLR

Questionnaire	Found in	Description
2DES	[127]	Computer program used to collect continuous reports on emotion provided by the study participants. Defined by dimensions valence and arousal [127].
Aesthetics scale	[127]	Measures aesthetic quality in websites [127].
Affect grid	[127]	A scale to assess affect on the axis of pleasure-displeasure and arousal-sleepiness [127]. The participant marks their current emotional state on the grid.
After Scenario Questionnaire (ASQ)	[55]	Single item questionnaire, so the response equals the overall result [55].
AttrakDiff	[42][91] [127]	Using this questionnaire, one can assess the user's feelings about a system. AttrakDiff studies hedonic and pragmatic dimensions of UX.
Attrack-Work Questionnaire	[127]	Based on AttrakDiff but elaborated for the context of mobile news. Can be filled in right after the participant used the system.
Differential Emotions Scale (DES)	[127]	A standardized instrument in the form of a checklist that divides someone's emotion experience into categories of emotion. Meant to get a sense of the emotional state of individuals at specific points in time [127].

EMO2	[127]	An instrument to measure emotion during the use of a product over time that provides feedback to designers. Extended data can be gathered through self-confrontation. Participants are filmed during their interaction with a product and afterwards immediately presented with the footage. They can report their feelings while watching this footage [127].
Emotion sampling device	[127]	ESD is a series of questions meant to find out the emotion the user is experiencing as the result of an event [127]. It asks about the causes of the emotion. ESD can identify 17 different emotions.
Game Experience Questionnaire	[127]	Can be applied after playing a game, multiple times over a longer period of time. Measures experiences during game play, gaming with others and after the user has stopped gaming [127].
Geneva Appraisal Questionnaire	[127]	Can be used to describe emotional experiences by assessing them through recall and verbal report [127].
Geneva emotion wheel	[127]	A method where users can self-report emotions that are brought about by certain events or objects [127]. The user chooses which emotion they feel based on an emotion wheel.
Hedonic Utility Scale (HED/UT)	[127]	Measures the attitude of the participant using 12 items to measure the hedonic value and 12 items to measure the utility or usability value of a product, service or concept [127]. This scale is subjective and sometimes shows some sensitivity issues.
Human computer trust	[127]	Human computer trust can be measured as the willingness of a user to act on the recommendations, actions and decisions of an AI decision aid and the degree to which a user is confident in these recommendations, actions and decisions [82]. This method consists of several scales measuring cognitive and affective components, where the affective components are the strongest indicators of trust [127]. The scales are subjective and this method requires functional prototypes or existing products.

Intrinsic Motivation Inventory (IMI)	[59][127]	The Intrinsic Motivation Inventory (IMI) is a subjective scale that measures multiple dimensions of a participant's experience of an activity in a lab setting. The scale provides subscale scores of the participant's effort, interest/enjoyment and value/usefulness, amongst other values [127]. From the subscales, the interest/enjoyment scale is considered the self-report measure of the IMI and often contains more items than other subscales.
IsoMetrics	[78]	Standard questionnaire, aims to offer usability data for summative and formative evaluation.
Mental effort	[127]	The Mental Effort Scale helps determine how much (perceived) mental effort was required to complete a task [127]. Should be used in addition to other methods to establish a broader picture.
MUSiC performance measurement method	[98]	MUSiC (Metrics for Usability Standards in Computing) performance measurement method is a method for deriving performance-based usability metrics [15]. To aid in video analysis, a software called DRUM can be used [81,98]. It provides measures of the effectiveness and efficiency of the system use, as it evaluates if and to what extent a task has been completed and how much time it took to complete a task. It also analyzes the amount of time that was spent unproductively (e.g. asking for help) [15]. Using this method, comparison with earlier prototypes or competing products is possible.
Net Promotor Score (NPS)	[55]	A type of product experience tracker (see table ??). Single item questionnaire, so the response equals the overall result [55].
PAD	[127]	Scale that can average responses of users to stimuli. PAD looks at pleasure (P), arousal (A) and dominance (D) scores [127]. It is a subjective scale, and the dominance scale is sometimes difficult to understand for users.
Positive and Negative Affect Scale (PANAS)	[127]	Originally developed for a clinical setting, the PANAS is a psychometric scale that measures the moods of users, so positive and negative affect, both as states and as traits [127]. This method has been heavily researched, validated and adapted (e.g. short form [118]).

Presence questionnaire	[127]	This questionnaire is designed to measure the presence a user feels, meaning the subjective experience of being in a certain environment (there) while physically being in another environment (here) [132]. The PQ is typically used to measure presence in games or virtual environments [127]. This subjective scale is well studied for use of games and virtual environments.
Product Attachment Scale	[127]	The product attachment scale can help quantify the attachment a consumer has to a product, which in this case is defined as the emotional bond a customer has with a durable product [89,110]. The scale consists of four items that can be rated on a Likert scale of seven points [127]. This method looks at the long term use of a product, but is subjective.
Product Emotion Measurement instrument (PrEmo)	[127]	The PrEmo is an instrument to assess emotional responses a consumer may have to a product [33]. It is a subjective self-reporting instrument and contains a set of 14 emotions portrayed by a virtual cartoon character displaying dynamic facial, bodily and vocal expressions. Participants are asked to select the animations that fit their responses. This instrument does not rely on language and could thus be used in a cross-cultural setting. Measures static stimuli (e.g. fragrance, taste, appearance), not dynamic stimuli (e.g. product usage) [127].
Product experience tracker	[127]	A survey is sent out to a subset of users after their real-world use of a product. This is done as soon as possible after use. The survey measures metrics such as likeliness to recommend or satisfaction of a product. The survey is sent out periodically or continuously to track whether the experience changes over time [127]. An example of this is the NPS or Net Promotor Score, which can be found in table ??.

QSA-GQM questionnaire	[127]	The QSA-GQM questionnaire aims to measure the intrinsic motivation a person has to acquire knowledge [127]. It is used in software engineering to review software quality. The QSA (Questionnaire on Software Assessment), a questionnaire on learning strategies, is combined with GQM. QSA measures the properties of software, difficulty of working with the software and the user's satisfaction with the software [90]. The QSA questionnaire is structured according to the GQM (Goal Question Metric) approach.
Reaction checklists	[127]	Summative method that allows for lightweight testing. After the participant has used the system, they are presented with a list of possible reactions to it. Can be applied to collect first impressions and initial responses to a product [127].
Self-assessment manikin (SAM)	[39][59] [127][134]	The self assessment manikin (SAM) is a technique that measures the pleasure, arousal and dominance that are associated with the user's affective reaction to stimuli [18]. SAM depicts cartoon characters representing pleasure, arousal and dominance. It is a quick, subjective non-verbal technique that can be applied in many contexts. The dominance scale can be difficult to understand [127]. This method is based on the PAD method. Relatively easy to apply and analyse [39].
Sensual evaluation instrument	[127]	A self-assessment tool of affect during human computer interaction /citeisbister2006sensual. While the user interacts with a system, they have access to physical objects with different shapes (e.g. spiky, smooth). When emotions arise during their interaction with the system, the participant is asked to pick one of the shapes to express the emotion [127]. This method is non-verbal and allows for studying ambiguity. It is not easy to analyze this method. The method is also not generalizable.
ServUX Questionnaire	[127]	Questionnaire for evaluating service UX. Consists of modules that each address a different aspect of ServUX [127]. To be used after the user has used the service.
Single Ease Question (SEQ)	[55][59]	Single item questionnaire, so the response equals the overall result [55].

Software metrics / usability metrics	[4][96] [97][98]	This method should aid in establishing observable, quantitative measurements. Usability metrics can quantify the usability of a system based on effectiveness, efficiency and satisfaction [4]. A representative number of users is required so the results can be generalized. Examples of usability metrics are the standardized scales SUS and the UMEX. These can be found in table ?? below.
Subjective Mental Effort Questionnaire (SMEQ)	[55]	Single item questionnaire, so the response equals the overall result [55].
SUMI	[59][78] [127]	A 50-item questionnaire to compare competing products and different versions of the same product. A minimum of 10 users is recommended. Requires user input through a keyboard, screen or pointing device [68,69]. Analyzes a product through the eyes of the end users [78].
Survey / questionnaire	[37][38] [42] [78][85] [96][97] [98][130] [137]	Representative users answer a list of questionnaire items according to a Likert scale [98]. The statements are intended to measure a certain usability aspect or dimension of the user's satisfaction. Many different questionnaires exist, the ones that were found in the MLR can be found in Table ?? below. It is also possible to create your own questionnaire, but there is a validity and reliability risk.
System Usability Scale (SUS)	[55][57] [59]	Can indicate whether or not there is a problem with the usability of a product, but cannot pinpoint the exact problem [41]. Single item questionnaire, so the response equals the overall result [55].
Usability Metric for User Experience	[41]	Four-item Likert scale to measure an application's perceived usability [41]. It is a subjective assessment designed to provide results similar to the SUS. This scale was introduced to fill in the gaps that the SUS left.
User Experience Questionnaire	[55][57] [91]	Subjective scale on user experience.
User model checklist	[78]	Based on user's cognition-motor chain in specific tasks. From the three found in the paper, this one was the best rated [78].

UTAUT	[59][127]	UTAUT (Unified Theory of Acceptance and Use of Technology) is based on Technology Acceptance Model (TAM) but addresses some shortcomings, including affective aspects [127]. This method is a tool to assess the likelihood of success when introducing new technology and can help understand the drivers of acceptance so that interventions like training or marketing can be introduced. UTAUT targets users that may be less inclined to use new systems [125]. It is subjective, but well studied and tested. Often used for website and computer applications.
Website Analysis and Measurement Inventory (WAMMI)	[127]	20 item questionnaire measuring UX based on reactions of users. It benchmarks the website relative to other websites in the WAMMI database and generates objective data. It analyses qualitative comments and reactions visitors have to your website and interprets qualitative and quantitative data to determine what areas to improve upon.

Table 17: Software tool methods from the MLR

Method	Found in	Description
Automated evaluation via software tool	[96][97] [98][117]	All activities required for a usability evaluation are performed using a software tool. Some tools are able to simulate human actions, others just perform metric-based measurements based on the user's activities [98]. These systems can also create a log file to be analyzed after testing.
Click map / scroll map / heat map	[49][98]	A click map (scroll map, heat map) is a visual representation of the attention of a user [98]. This can give insight in the most or least popular sections and elements users mistake as links. Can be obtained through software tool.
Context-aware ESM	[127]	The system detects the current context (e.g. time, location, nearby devices) and when predefined criteria are fulfilled, prompts the participant to record their experience [127]. Many different kinds of data can be reported, and in many different kinds of ways (e.g. questionnaire, text, audio recording, image etc.). Data can be sent to researchers immediately or stored for later analysis.

Emoscope	[127]	Focuses on the difference between what a user says they do and what they actually do through the Emotron, Emotracking and Pulsetron [127]. Emotron is a software that allows the collection of emotional data during a task and generates diagrams of task. Emotracker uses eye tracking to find the points of attention and creates a thermal map. The results of the Emotron and Emotracker are combined. Pulsetron collects physiological emotional data.
Eye tracking	[37][38] [59][85] [98][130]	Using eye tracking, a user's visual attention on a display can be captured using software or hardware (head mounted system, small camera) designed for this purpose [36]. This method can be enhanced with different sensors to allow for a more detailed analysis. By analysing the visual path of the end user, one can find out what the relevant information is, what sections are ignored and what information is overlooked [98].
Facereader	[127]	FaceReader can be used to track a user's affective state while using products or software. It analyses facial expressions from a video in real-time. After constructing a model of the face, the software calculates the likeliness of each of six basic emotions (joy, sadness, anger, surprise, fear and disgust) [127].
Feeltrace	[127]	A software tool to allow observers to self-report the perceived emotional content of a stimulus over time.
Field observation / field study / observation	[13][39] [60][85] [96][98] [127]	The researcher observes the user in their 'natural habitat', i.e. the field, in a direct or indirect way [98]. The researcher observes, takes notes and asks questions.
Field study: 3E (Expressing Experiences and Emotions)	[39] [107] [127][134]	During a field study, 3E can be used to collect information about users' experiences and emotions. This method is semi-structured /citevermeeren2010user. Useful for uncovering pragmatic problems [39].

iScale	[127]	iScale is a survey tool using the UX curve pen-and-paper method. The participant is asked to draw one or more curves to describe how their product experience changed over time [127]. The curve is drawn on a template consisting of a time line and a horizontal line to separate the positive and negative sides of the experience. This method is retrospective, meaning it relies on memories rather than reality, but it can help reveal the experiences the participant found most useful.
Kansei engineering software	[127]	This method concerns the design of affective values in product solutions [113]. It 'measures feelings and impressions and shows the correlation to properties of the product [127]. The Kansei software is a tool that can automatically collect data and evaluate this data according to the rules of Kansei engineering, thus making this method more efficient and easier to apply.
Opinion mining / sentiment analysis	[98]	Opinion mining is a method for extracting, classifying, understanding and assessing opinions expressed in user-generated content such as comments on social media [23]. This is done with the use of natural language processing and text analysis [98]. Opinion mining may also be called sentiment analysis [19].
Outdoor Play Observation Scheme (OPOS)	[127]	OPOS is an observation scheme that can be used to evaluate outdoor pervasive games that were intended for children [11]. It is an objective method that can help evaluators compare pervasive games based on the play behaviours they bring. The objective measurements can be combined with the subjective opinions of the children after the interviews [127]. Coding the videos may be very time consuming.
Product reviews	[24]	Using this method, the user experience of a product can be evaluated based on its online reviews. The reviews objectively show the user's opinion and contain a large amount of data [24]. An advantage of this method is that participants do not need to be selected, as the product reviews already exist. However, not all buyers leave a review, so the user's motivation for leaving a product review should be taken into account.

TRUE (Tracking Real-time User Experience)	[127]	While playing a game, participants' behaviour and reactions are recorded on logs and video. This method is objective and rapid and can be done in groups. It can be used to assess design goals, too. However, this method does require software development and a lab [67,127].
TUMCAT	[127]	Users install a software package that logs their actions and sends these loggings to a server to be studied remotely [127]. Offers context, can be done remotely and long-term measurement is possible. Software is needed and triggers should be defined [71,126].
User's feedback	[49][137]	Any type of feedback from the user. Seyff, Ollmann and Bortenschlager [115] describe AppEcho, which is a mobile feedback approach where individual user feedback is collected in situ. The feedback can then be used for new requirements. Other (mobile) applications for the same purpose exist.

Table 18: Workshop methods from the MLR

Method	Found in	Description
Canvas card sorting	[98]	Similar concept to card sorting, where users are asked to select the most valuable concepts and arrange them according to a template. Main categories are pre-established [98].
Card sorting	[98]	Card sorting can be used to verify the way information is structured in a product. Required paper cards with a word or a phrase written on one side that represent a concept that is part of the user interface. Participants are given a stack of cards and are asked to group them as it makes sense to them. If a taxonomy becomes visible, this can be used in the UI [98].
Co-discovery	[98]	A pair of users (friends) explore a product or concept together. This can be with a researcher present to guide the discussion or using a video tape to record it. The friendly and familiar setting typically triggers more experiential comments than discussing it with a researcher [127].
Emocards / emofaces / emotion cards	[107][127]	Using emotion cards, users can quickly visually document their emotions at a specific moment [127,32].
Experience recollection method	[75]	Users are asked to remind their past experience and rate their degree of satisfaction on a scale of -10 to +10. This is based on the UX curve [75].

Expert walkthrough (group-based)	[127]	A group of domain experts evaluate a user interface to identify usability-problems, possible design improvements and successful design solutions [127]. The evaluators do not need training. With the combination of probing, UX problems can also be identified.
Focus group	[13][35] [38][83] [96][97] [98][130] [137]	A representative group of users participate in an open discussion where they are free to talk and listen to each other, analyzing the graphical interface of a software product [98]. Users can develop their own ideas. Can give insight into people's thoughts and motivations.
Fun toolkit	[127]	Again-Again table and a Fun Sorter. These are used to measure three dimensions of fun: expectations, engagement and durability [127]. Designed for children.
Laddering	[127]	Laddering in UX can help the researcher understand in what way product attributes benefit end users' personal values [2]. Using this method, more in-depth information can be elicited compared to pictorial representational scales (e.g. SAM) [136]. It shines light on the reason users like or dislike a product. Laddering is similar to think-aloud, but is a more gated and incremental procedure. This means this method would work well when working with groups like younger children.
MAX	[20][39]	The Method for the Assessment of eXperience (MAX) uses cards and a board to motivate users to report their experience [20]. This method does not require experienced UX researchers (i.e. can be used by software engineers) and is meant to evaluate the user experience of finished or prototyped software products. MAX can capture the UX of a product quickly and provides a structured way to conduct a UX evaluation with the space for the participants to explain their reasoning. Relatively easy to apply and analyse [39].

Mindmap	[127]	A mind map or concept map is a practical tool that aids to visualise the thinking process [129]. In a mind map, a non-linear network of connected and related concepts is shown and creating a mind map requires spontaneous, free-form thinking to find creative associations between concepts [28]. In the concept of UXEMs, a mind map can be used to show the creator's mental model of a system to identify misconceptions and problematic areas [47]. This method can be applied to study longer periods of time.
Multiple sorting method	[127]	Multiple sorting method is a variation of the Repertory Grid Technique where multiple designs are shown to the participant, who is then asked to group the designs in as many categories or piles as they desire. They are then interviewed about their motivations for doing so [127]. After, different design triads are shown until the participant cannot group the designs in new ways. Then, as multidimensional scalogram analysis can be conducted to construct spacial maps. These maps can be analysed with the interview data.
Paired comparison / pairwise comparison	[127]	Pair combinations of stimuli are made and presented to the participant, who is then asked to select the 'best' of the pair [127]. The data from this forced choice of all participants is pooled together, after which an ordering of the stimuli set can be created. This method is suitable to use with children and is generally easy to use.
Pencil & paper	[98]	Users evaluate aspects of a system prototype on paper, where they may modify the interface design. They may also comment and annotate their observations [98].
Perspective-based (usability) inspection	[38][98] [127][130]	With perspective-based (usability) inspection, every inspection session the researcher focuses on a different subset of usability issues based on a usability perspective [138]. This method should uncover more problems than general inspection methods because of the combination of different perspectives. Usability perspectives may include aesthetics, fun, comfort and other user experience [127].

Product personality assignment	[127]	In product personality assignment, the participants are presented with a list of characteristics or personalities based on Myers-Briggs type indicators, e.g. friendly, sensible. The participants are asked to assign these personalities to product designs and motivate their reason for doing so [127].
Prototype evaluation	[96][97] [98][137]	An end user and usability expert participate in a meeting where the user is asked about their expectations for a mockup or prototype of a concept, product, service or idea [98]. The main purpose for prototypes is to test purposes or find and eliminate bugs [103].
Repertory grid technique (RGT)	[127]	The repertory grid technique (RGT) is used to systematically explore the views a participant has on a certain topic [48]. This is done based on constructs, both emotional (e.g. warm-cold) and rational (e.g. professional-popular). These constructs are based on the personal construct theory [65], which states that everyone makes sense of the world through their own personal bipolar constructs. The extraction of these constructs can be done by presenting the user with a set of three artifacts. The user is then asked to describe how two artifacts are similar and how the third is different from them [53]. This is repeated until no new constructs are found. This technique is structured but still dynamic and open to the personal constructs of a user. This method exists on the border between qualitative and quantitative [127]. RGT does require a significant amount of effort, both by the researcher and the user. It is less time-consuming when compared to fully open approaches (e.g. unstructured interviews).
Resonance testing	[127]	Resonance testing is a qualitative method primarily intended for use with physical products [127]. Using resonance testing, the researchers can test product concepts for their emotional and functional design attributes [80]. Multiple abstraction levels are required, meaning artifacts at different fidelity levels are needed. This is a straightforward method that has been proven in practice, but is expensive.

Semiotic inspection method	[98]	Semiotic inspection examines the diversity of signs (e.g. widgets, words, colours, images, graphic layouts) to which users are exposed as they interact with a system [30]. This qualitative method is subjective and the quality of the data is dependent on the expertise of the researcher. This method may be used to complement a communicability evaluation. Semiotic inspection focuses on the elements within a system or interface, not the way the user interacts with the system [98].
Sentence completion	[91][127]	Sentence completion takes place after the user has used the system [127]. The beginnings of sentences are presented to the user to trigger them to think of different aspects of product use. This qualitative method yields structured data about users' views [74]. The results are more time consuming to analyze compared to other qualitative methods, but can retrieve more information on a user's negative feelings. Compared to traditional questionnaires, the data from sentence completion is more reliable as users can express themselves more freely [127]. Sentence completion can also be used online to reach a broader user group. It may be used to complement other methods as well.
Simplified pluralistic walkthrough	[98]	A simplified pluralistic walkthrough typically takes place after a simplified streamlined cognitive walkthrough [50]. SPWs usually take place in two sessions with four customers each. Real product users have the spotlight in these sessions to gather real-world feedback on the usability of the features. [50] created the SPW as a step towards informal cognitive walkthroughs, where SPWs are combined with simplified streamlined cognitive walkthroughs.

Simplified streamlined cognitive walkthrough	[98]	A simplified version of a streamlined cognitive walkthrough, where the workload and responsibilities for the researcher are increased to decrease the amount of time required for training the participants [50]. For each step, the researcher answers two questions: (1) as the user, would I know what to do at this step? (2) If I do the right thing, as the user, do I know I have made progress toward my goal? After these questions are answered, the participants are asked to share their opinions. Documentation from this session exists as notes taken during the session, which include the changes to be made after this session. [50] created the SSCW as a step towards informal cognitive walkthroughs, where SSCWs are combined with simplified pluralistic walkthroughs.
This-or-that	[127]	Typically aimed at children [127]. Two conditions (systems or versions) are presented. After a while, the participant is asked to choose: this or that?
Usability & communicability evaluation method	[98]	Assesses the communication and determines to what degree the designers succeeded in conveying their intents and interactive principles through the interface [95,101]. While a representative number of users interact with the product, evaluators identify communication breakdowns. Categorization is done with the use of thirteen expressions or labels. The evaluator interprets the issues and rebuilds the message to identify possible improvements [98].
User testing	[37][38] [83][96] [97][98] [117][130]	A representative number of end users follow a predefined list of tasks while interacting with the software, or are asked to explore it freely. Through observation of these tasks, usability and design issues can be identified. Commonly applied in a usability lab where a user's gestures and screen can be closely monitored [98]. There is no consensus regarding the number of users required. While earlier research [128] states that 85% of usability problems can be found with four or five participants, others state that five participants only identified 35% of the usability problems [116]. The number of participant required ultimately seems to depend on the type of user test [6].

User testing - log analysis	[38][130]	Usage data is analyzed by testers or software tools. It is typically used post-release to collect data regarding the field of use of a system [56]. It can also be used to collect more detailed data during user testing. When related to gaze points this is categorized as eye tracking.
User testing - performance measurement	[38][130]	Usage data is recorded by testers or software tools. Statistics are obtained during the test. Performance measurement may be formative or summative. This method can be applied to obtain a more thorough understanding of a user's need and to improve the product to provide better user experience [14].
User testing - question asking	[38][130]	Testers ask the users direct questions. This can be seen as an alternative to the thinking-aloud method [64]. This method aids in identifying the problems users can experience in different contexts, what information a user may need, what features are difficult to learn and how users could misunderstand the system.
User testing - remote testing	[37][38][130]	Testers and users are not located in the same place during the user test. The usability issues that can be identified through this method are similar when compared to a lab setting [119]. Often performed at the same time as log analysis methods. Tullis et al. [119] describe two different types of remote testing: synchronous and asynchronous.
User testing - retrospective thinking aloud	[98][22]	Very similar to thinking aloud, only with this method the user is asked to formulate their thoughts retrospectively, so after the user testing session. Recordings of the session may be used to aid the narration of the user [98]. In this method, usability problems may be identified through verbalisation by the user, as opposed to observation by the researcher [124].
User testing - thinking aloud / thinking out loud	[37][38][39][58][60][83][97][98][130][137]	User interacts with the system while verbalizing their thoughts. Supervisors encourage the user to speak their mind and give their opinion on the system. This is a way for researchers to gain insight into the thought processes of a user [17]. It seems like thinking aloud while the user is performing the actions outperforms retrospective thinking aloud or a hybrid method, not only on the numbers of usability problems identified but also on the time required on the researcher's part [5]. Useful for uncovering pragmatic problems [39].

User testing - extended usability testing	[37][38] [83][96] [97][98] [117][130]	Representative amount of end users follow a pre-defined list of tasks while interacting with the software, or are asked to explore it freely. Through observation of these tasks, usability and design issues can be identified. Commonly applied in a usability lab where a user's gestures and screen can be closely monitored [98]. There is no consensus regarding the amount of users required. While earlier research [128] states that 85% of usability problems can be found with four or five participants, more recent research states five participants only identified 35% of the usability problems [116]. The amount of participant required ultimately seems to depend on the type of user test [6].
UX curve	[85][127]	UX Curve method aims at assisting users in retrospectively reporting how and why their experience with a product has changed over time. A method of retrospective reporting. Users are aided by the UX curve to report how and why their experiences with a product changed over time [127] [73]. According to Vermeeren et al. [127] this method can help identify the most impactful experiences over time, but it is less reliable as it counts on memories rather than reality. Is a pen-and-paper method, but iScale is a related tool.
Valence method	[127]	Based on Hassenzahls model where UX is defined as a primary evaluative feeling while using a product of service [127]. Users are asked to continuously report all positive or negative feelings using a remote control (e.g. green checkmark for positive, red x-mark for negative). Short sessions (<10 minutes). Afterwards, a retrospective interview is conducted to investigate for each valence marker what aspect caused the feeling and what the personal meaning and underlying needs are.

Workshops + probe in- terviews	[49][127]	Design probes are a tool to help understand human phenomena and explore design opportunities. They are based on user participation through self-reporting, take user's personal context and perceptions into consideration and have an exploratory character [87]. After the initial exploratory user research using probes, the same participants are invited back to validate the analysis in an interview. In this session, participants can experience the early prototypes and provide feedback. This is done in a group setting citevermeeren2010user.
-----------------------------------	-----------	--

B Informed consent

Thank you for your interest in this study! This study aims to create a decision model to aid in the selection of UX evaluation methods. Your responses will be kept strictly confidential, and digital data is stored securely. Any publications based on this research will not include your name or the company name. In any publications within the company, your name will be anonymized, but the location of your team may be included. The project's research records may be reviewed by departments at Utrecht University responsible for regulatory and research oversight. Your participation in this research is voluntary. You may choose to withdraw your participation at any time without being penalized.

If you have any questions, comments or concerns about this study, you may contact a.e.stolk@students.uu.nl or thesis supervisor c.vannimwegen@uu.nl.

C Interview protocol: Current situation and methods

C.1 General

- Welcome interviewee.
- Introduce researcher and research.
- Ask to read and sign informed consent.

C.2 UXEMs

- Explain UX evaluation methods and how they differ from UX design methods.
- (Q1) Off the top of your head, what UX evaluation methods do you usually use?
- Show list of UXEMs compiled from earlier interviews. Compare it to the answers given to Q6.
- (Q2) Looking at this list, do you see any UX evaluation methods you may not have mentioned yet?
- Show list of UXEMs from the literature.
- (Q3) Looking at this list, do you see any UX evaluation methods you may use?
- (Q4) Or ones you have heard of, but do not use? Why do you think that is?

C.3 Current process

- I would like to ask you some questions about your typical UX projects. Imagine a new project is about to start.
- (Q5) How do projects typically get assigned?
- (Q6) Who decides what to do?
- (Q7) How do you pick a UX evaluation method (or multiple) for your project? What do you base your decision on?
- (Q8) If a colleague from the same team and location would get this project assigned, would they handle it the same way? Why?
- (Q9) If a colleague from another UX team and another location would get this project assigned, would they handle it the same way? Why?

C.4 End

- (Q10) Do you have anything you would like to add?
- Thank participant for their time and efforts.