

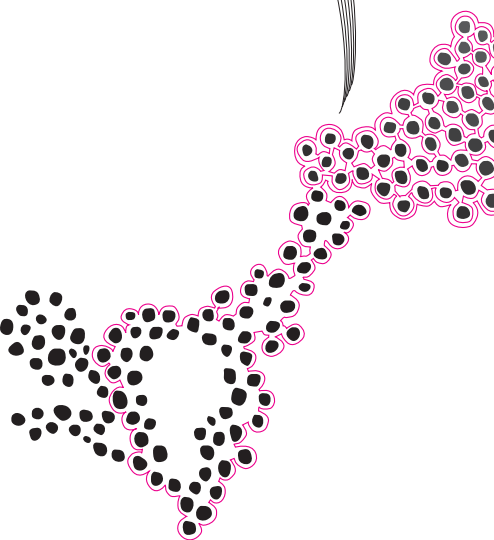
MSc Interaction Technology  
Final Project report

# AI assistance for UX testing: Exploring its impact on the UX researcher



Siauw Ying Liem

Supervisor: Mariët Theune, Anis Hasliza Abu Hashim, Maro  
Gómez Maureira, Berend-Jan van der Zwaag & Marloes Aben



April, 2024

Department of Interaction Technology  
Faculty of Electrical Engineering,  
Mathematics and Computer Science,  
University of Twente

## **Abstract**

This thesis explores the integration of artificial intelligence (AI) assistance in user experience (UX) testing workflows, focusing on its impact on UX researchers' behaviour, experiences, and cognitive load. Through experiments conducted in collaboration with a Dutch UX research agency, the study investigates the specific effects of AI assistance on tasks such as logging observations and discussing findings during simulated test days. Results indicate that the extent of the impact of AI assistance varies per task, but also among participants, influenced by factors such as perceived usefulness in terms of efficiency and effectiveness, and ethical considerations. The decision to incorporate AI assistance into UX workflows requires careful consideration of functionality, trust in AI output, and potential ethical risks.

*Keywords:* AI assistance, UX research, UX testing, LLM, AI textual data filtering assistance, cognitive load

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Related literature</b>	<b>3</b>
2.1	Artificial Intelligence	3
2.1.1	Definition of AI	3
2.1.2	Qualitative coding	5
2.1.3	Summarisation	10
2.1.4	Skimming	11
2.1.5	Takeaways for AI textual data filtering assistance	12
2.2	Cognitive Load	13
2.2.1	What is cognitive load?	13
2.2.2	Types of cognitive load	13
2.2.3	Factors for cognitive load	14
2.2.4	Measuring cognitive load	14
<b>3</b>	<b>UX testing workflow</b>	<b>17</b>
3.1	Test day	18
3.1.1	Obslogging	19
3.1.2	Debriefing	22
3.2	Report-making day	23
3.2.1	Qualitative analysis	23
3.2.2	Report making	24
<b>4</b>	<b>Methodology</b>	<b>26</b>
4.1	Participants	26
4.2	Variable Manipulation & Measurement	27
4.2.1	Independent variable: AI textual data filtering assistance	27
4.2.2	Dependent variable for RQ1 & RQ2	29
4.2.3	Dependent variable for RQ3: ECL & GCL	30
4.2.4	Controlled variables	30
4.3	Test day simulation	31
4.4	Materials	31
4.5	Data analysis	32
4.5.1	Data collection	32
4.5.2	Statistical analysis	32
4.5.3	Qualitative analysis	32
<b>5</b>	<b>Experimental procedures</b>	<b>33</b>

5.1	Pilot study	33
5.2	Experiment	33
<b>6</b>	<b>Qualitative results</b>	<b>36</b>
6.1	Obslog assistance	37
6.1.1	View of AI obslog assistance	37
6.1.2	Usage of AI assistance	37
6.1.3	Experience with the AI assistance	38
6.2	Debrief assistance	40
6.2.1	View of AI debrief assistance	40
6.2.2	Usage of AI Debrief assistance	42
6.2.3	Experience with the AI assistance	48
6.3	Experimental set-up	49
6.4	Obslog vs Debrief AI assistance	50
<b>7</b>	<b>Cognitive load results</b>	<b>52</b>
7.1	Obslog task	53
7.1.1	Mental effort (Paas scale)	54
7.1.2	ECL	54
7.1.3	GCL	54
7.2	Debrief task	54
7.2.1	Mental effort (Paas scale)	56
7.2.2	ECL	56
7.2.3	GCL	56
<b>8</b>	<b>Discussion</b>	<b>57</b>
8.1	Obslog assistance	57
8.1.1	Usage of AI assistance	58
8.1.2	Experience with AI assistance	59
8.1.3	Cognitive Load	62
8.2	Debrief assistance	63
8.2.1	Usage of AI assistance	63
8.2.2	Experience with AI assistance	66
8.2.3	Cognitive Load	66
8.3	Obslog & Debrief assistance	67
8.4	Limitations	68
8.5	Future works	69
<b>9</b>	<b>Conclusion</b>	<b>71</b>
	<b>Appendices</b>	<b>75</b>
	<b>A Interview testscript</b>	<b>76</b>
	<b>B Cognitive Load Questionnaire</b>	<b>84</b>
	<b>C Transcript translated excerpts for Results</b>	<b>89</b>
C.1	Obslog	89
C.1.1	Usage	89
C.1.2	Experience	89
C.2	Debrief	91

C.2.1	View of AI debrief assistance . . . . .	91
C.2.2	Usage of AI debrief assistance . . . . .	92
C.2.3	Evaluation of AI output . . . . .	96
C.2.4	AI references . . . . .	98
C.2.5	Ethical risks . . . . .	98
C.2.6	Experience of AI debrief assistance . . . . .	99
C.3	Experimental set-up . . . . .	100
C.4	Obslog vs Debrief . . . . .	101

# Chapter 1

## Introduction

In recent years, the accessibility of artificial intelligence (AI), particularly generative AI like ChatGPT, has sparked widespread interest and curiosity (Faber, 2023). This surge in attention has led to many companies being eager to harness its capabilities to their benefit. According to a generative AI executive survey conducted by the Capgemini Research Institute with  $n = 800$  organisations, 96% of the companies have had board meetings to discuss the use of generative AI (Engels, 2023). Popular sentiment is that integrating AI in their systems and work procedures has the potential to improve productivity, ease of task, costs, etc. The same goes for the field of UX research, which is short for user experience research. UX research is a broad field with many research methods, of which the approach can differ between companies and UX researchers. There are likely endless options on how to exactly integrate generative AI into one's workflow, in terms of functionality but also interaction and interfacing. Rather than investigating the best way to apply and employ AI assistance, we are more interested in exploring the possible general implications and impact of AI on the UX workflow. For example, what effect will the use of AI assistance have on the UX researchers and how they approach their tasks? What thoughts, feelings, impressions, and behaviours will they have concerning the AI assistance and their interaction with it? Can AI assistance influence the difficulty of the tasks or how much mental effort is expended? Mental effort is an aspect of *cognitive load*, which will be described in more detail in Ch2. What can those findings tell us for future employment and integration of AI within UX workflows?

This research will investigate the impact of AI on the UX workflow on a use-case basis in collaboration with a Dutch user experience (UX) research and strategy agency, which provides insights into contemporary UX research practices. This agency will from here on be called the thesis company.

To narrow the scope, this thesis is only focused on the workflow of UX testing projects, which is the most used method at the thesis company. More specifically, we look at the tasks of obslogging and debriefing on the actual test day, which will be explained in Chapter 3. Moreover, we will focus on AI assistance (also coined AI augmentation or collaboration) rather than complete automation, as we are more interested in the potential of combining AI and human input/output. This leads to the following research question for this thesis:

***“How and to what extent does AI assistance have an impact on the UX testing workflow?”***

To answer this question, we look at the impact from several angles, namely UX researchers'

behaviour, experiences with AI assistance and cognitive load during the UX testing workflow. This is formulated in the following subquestions:

RQ1: *“How does AI assistance influence how UX researchers perform the tasks of noting down observations and debriefing?”*

RQ2: *“How do UX researchers experience the AI assistance during a UX testing test day?”*

RQ3: *“To what extent can AI assistance reduce the cognitive load during a UX testing test day?”*

RQ1 concerns UX researchers’ behaviour, where behaviour refers to their usual actions and approach to performing tasks versus with AI assistance. This includes their use of and interaction with their usual tools and the integrated AI assistance.

For RQ2, we define experience as the UX researchers’ thoughts, feelings and impressions when they go through a specific interaction, be it with their standard tool or with the AI assistance (“What is User Experience?”, 2020; “What’s the difference between CX and UX?”, 2020). Experience is a broad term, so to somewhat constrict the possible deluge of findings, we focus on the more binary values of positive and negative experience, also in relation to the desire for and openness to future usage.

Regarding RQ3, what cognitive load is will be explained in Section 2.2. Why this research focuses on the metric of cognitive load will become clear in Chapter 3. What AI (assistance) entails in the context of this thesis will be outlined in Section 2.1; the specifics of the implemented AI system for the experiments are described in Chapter 4.

In this thesis, we will first explore related literature on AI used for UX research or similar fields and tasks, and cognitive load, see Chapter 2. Next, we provide information on the thesis company’s UX testing workflow, including its exact procedures, utilised tools and challenges in Chapter 3. Chapter 4 describes the methodology used for the experiments, and Chapter 5 outlines the experimental procedures. Then Chapters 6 and 7 show all the findings, both qualitative and quantitative. Lastly, the results and limitations are discussed in Chapter 8, before providing the conclusion to this research in Chapter 9.

## Chapter 2

# Related literature

In this chapter, we explore AI technologies and tools that can be employed as assistance for the UX testing workflow. We examine relevant literature to identify suitable options and important considerations. Next, to be able to answer RQ3, we briefly examine the literature on cognitive load, including what it entails, its factors, and how to measure it.

### 2.1 Artificial Intelligence

Before we dive into the vast scientific research space of AI, we first further narrow down the scope. UX researchers conduct many tasks that (mostly) involve a lot of textual data handling, which becomes apparent in the next chapter (Ch3). To cope with the large amounts of data, UX researchers (try to) apply filtering strategies. To find more suitable technology and tools for the UX testing workflow, this chapter examines related works that involve AI technology that could help with *textual data filtering*.

Within the literature on assistive AI technology, several domains and application areas are relevant for textual data filtering, namely qualitative coding (QC), summarisation and skimming. The current AI-assisted tools and technology for these topics will be further discussed in this chapter. However, we first give our definition of AI and outline relevant forms.

#### 2.1.1 Definition of AI

Artificial intelligence can be defined as “[The automation of] activities that we associate with human thinking, activities such as decision-making, problem-solving, learning, etc.” (Bellman, 1978). As the definition demonstrates, AI encompasses a vast field. For this thesis, we will only further detail the relevant concepts of machine learning (ML) and natural language processing (NLP), which are capabilities deemed necessary to achieve the aforementioned definition of AI.

Machine learning concerns the computer model having the ability to adapt to new circumstances and to detect and extrapolate patterns, in other words being able to “learn” (Russell, 2010). Within AI, there are three main types of learning, utilising different feedback strategies. The types are *supervised*, *unsupervised* and *reinforcement* learning.

For supervised ML, the model is provided with data consisting of input-output pairs from which it learns the patterns (a function) that can map the input to the output. An often applied form of supervised learning is classification (also often called detection), where



the output values are labels describing the input data (Samoili et al., 2021). Such labels are often manually created by humans. For adequate model performance, considerable amounts of labelled training data are needed.

For unsupervised ML, no feedback is given, instead, the model finds underlying patterns in datasets without pre-existing labels. A well-known form of unsupervised learning is clustering, where similar data is grouped together into clusters.

Lastly, reinforcement learning refers to models that learn through trial and error, getting feedback on decisions made in the form of rewards or penalties. Models of this type aim to maximise their rewards.

There exist many types of ML models; one such variety is called generative AI, which refers to models that generate new, synthetic data, resembling real (input) data.

The field of natural language processing (NLP) can also be regarded as a subfield of AI. NLP is, as its name suggests, a field concerned with the processing, analysis and synthesis of natural language and speech. In essence, it's about giving the computer the ability to communicate, using human language (Samoili et al., 2021). Some examples of NLP are speech recognition, chatbots and automatic text summarization. NLP draws from linguistics, finding patterns in language features, syntax, grammar, etc.

A typical NLP technique is term extraction or language feature extraction, which includes tokenisation, stop word removal, part-of-speech tagging (PoS; refers to the grammatical classification of e.g. verbs and nouns), stemming, lemmatisation, etc. Such features can be utilised to e.g. create NLP rules that can help capture language patterns.

Another technique within the field of NLP is large language models (LLMs), which are models that predict the probability distribution of language expressions (Russell, 2010). LLMs are generally trained using unsupervised learning, hence the model finds patterns within the textual input data by itself. LLMs can be used to predict the next words or phrases, based on provided input data. Nevertheless, natural languages are (semantically) ambiguous and constantly changing, hence language models will always be an approximation, using probability distributions of the possible meanings within language.

A technology that is currently well-known, called 'ChatGPT', combines ML and LLMs in a generative pre-trained transformer model (GPT; the latest available version at the time GPT3.5 turbo was used for this thesis). The model was trained using unsupervised learning on an enormous dataset, which includes scraped textual data from the internet (till 2021) and dialogue corpora. This process is often called the pre-training. Next, the GPT model was fine-tuned using supervised learning and reinforcement learning with human feedback to improve the model's performance for understanding and alignment with user intent, and for specific tasks, such as text summarisation and question answering. The model can be instructed using textual prompts, which it takes as new input (instructions). Combining it with a chat interface produces the ChatGPT application. (Gewirtz, 2023; OpenAI, 2017, 2022, 2023a, 2023b, 2023c)

The mentioned AI techniques and approaches are employed for the AI tools and technology described in this chapter. Each approach has its own benefits and limitations; e.g. supervised and NLP rule-based learning is generally more rigid in approach compared to unsupervised learning and is more time-consuming to train, although unsupervised learning can produce too unconstrained outputs.

### 2.1.2 Qualitative coding

An application domain of AI-augmented support systems that could be relevant for UX researchers and UX testing is qualitative coding (QC). QC is a process that is frequently used for qualitative analysis (QA) (Boyatzis, n.d.; Braun & Clarke, 2006), which can be used to analyse textual data e.g. interview transcripts. QC can be seen as a form of data filtering since it involves systematically organising and categorising qualitative data, producing a condensed and less complex version of the original data, allowing for easier analysis and interpretation. With QC, the researcher iteratively assigns labels, usually called codes, of varying levels of information to transcript segments (e.g. sentences or paragraphs) (Boyatzis, n.d.; N.-C. Chen et al., 2016; Jiang et al., 2021; Rietz & Maedche, 2021). During the UX testing workflow, similar processes are executed by adding labels with different levels of specificity to each observation and quote written down, for easier filtering at a later moment (see Ch3). The codes evolve over time, helping the UX researcher identify repeated patterns and discover emerging themes. This helps the UX researcher with making sense of the data and aiding in answering specific questions by e.g. filtering by theme (Braun & Clarke, 2006; N.-C. Chen et al., 2016). Coders often examine every data segment at least once to multiple times, making it a time- and effort-consuming task. Therefore, the objective of AI-assisted systems for QC is to reduce time and effort during analysis. How they achieve this differs per approach.

In the past few years, there seem to be three main approaches to AI-driven QC systems, namely rule-based, supervised ML-based (and often combinations of the two) and unsupervised LLM-based. An example of each approach will be discussed in this section.

The systems display the segmented raw textual data in some form and allow the UX researcher to freely code these segments. The assistance of the systems generally comes in the form of code suggestions for unlabelled text segments based on codes previously provided by the UX researcher or completely new code suggestions (LLM-based). Depending on the approach, user input (of the UX researcher) is utilised to different extents to generate AI output. The mentioned papers all create QC-support applications that maintain the context of the original data and reduce the to-be-analysed data to an extent.

#### Rule- & Supervised ML-based

The ‘PaTAT’ system from Gebreegziabher et al. (2023) is a rule-based system, where new codes are devised based on detected code rules in the user-inputted codes and text segments. These rules are composed of different NLP rules, such as lemmatised words, part of speech tags (e.g. verb, noun, adjective, etc.) or entity type tags (e.g. location, which will match any phrases corresponding with a location, such as Amsterdam) and soft matches (e.g. pricey will match synonyms like expensive and costly). An example of a PaTAT code rule for the code, price, is

“*MONEY + NUM + \* + NOUN*”

where NUM means number and a \* represents a ‘wildcard’, meaning it will match any sequence of words. The system generates numerous rules that are often variations of each other, e.g. different combinations of synonyms. PaTAT selects a small set of rules with minimal overlap to capture the most annotations. These rules are used to find unlabelled segments that match any of the code rules. The segments are then labelled with the corresponding code, displayed below the segment as a suggestion (see Fig. 2.1). The PaTAT system supports some user input by allowing the UX researcher to review, remove and update the code rules. For instance, they can directly edit what synonyms are allowed

to match.

Supervised learning is utilised to produce a confidence score for the code rules, “predicting” how well a specific code rule can predict a code. However, no supervised learning seems to be applied to produce the desired AI output, namely the code suggestions.

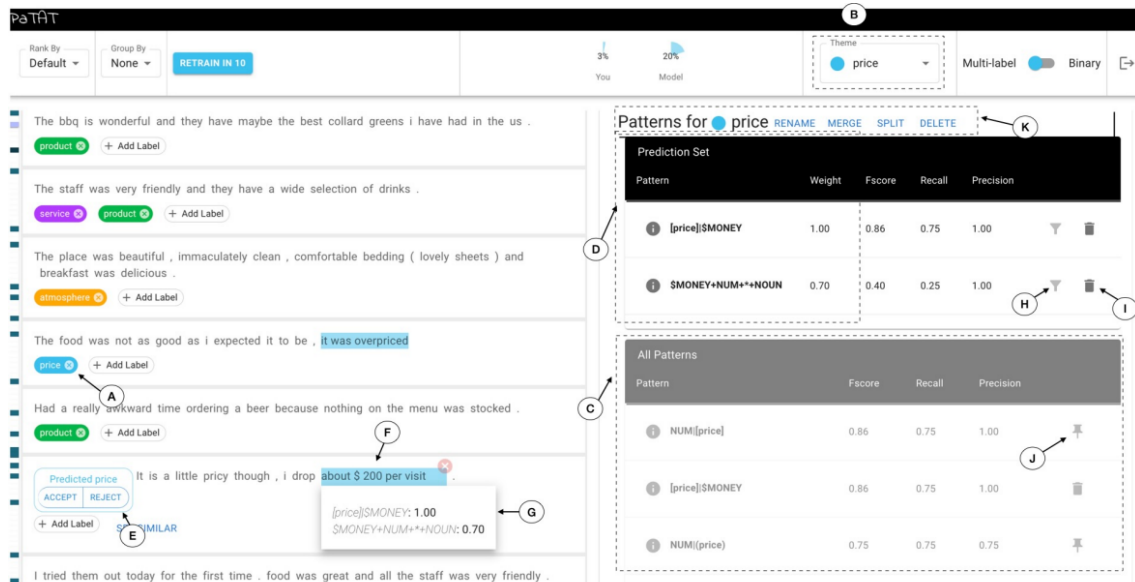


FIGURE 2.1: Gebreegiabher et al. (2023)’s PaTAT system

An AI-augmented QC system that combines the use of NLP rules and supervised learning to generate code suggestions is ‘Cody’ (Rietz & Maedche, 2021). Cody has a smaller syntax of rules than PaTAT, only supporting the use of complete or lemmatised words, boolean operators (AND and OR) and wildcard characters. Potential synonyms are identified by computing the similarity between the lemmatised text segment and each word in the code provided by the UX researcher. An example of a code rule in the Cody system for the text segment “Promotion not important” is

*“promot \* AND not AND (importan \* OR care\*)”*

where care\* is a potential synonym for importan\*. Cody’s rule-based code suggestions are produced the same way as PaTAT did. For the ML-based suggestions, Cody uses a supervised learning model to classify unseen text segments (model input) on the available user-inputted codes (ground truth labels). The predicted codes are displayed next to the colour-coded highlighted text as suggestions, along with a confidence percentage (see Figure 2.2). The model is continuously retrained after  $x$  number of edits and additions in real-time. Moreover, an explanation is given in the form of what keywords the text contained that had a large weight/impact on the prediction model. This is done by iteratively predicting a code while removing one to two words from the input and evaluating the accuracy. Users can accept or reject the suggested codes.

For QC, the more rigid approaches of rule- and supervised ML-based assistance have several limitations. For rule-based code suggestions, the AI assistance is restricted by the modelled syntax. Although incorporating more NLP rules allows the generation of code suggestions to be more flexible, the rich and unbound scope of human language remains difficult to capture to satisfaction using set rules.

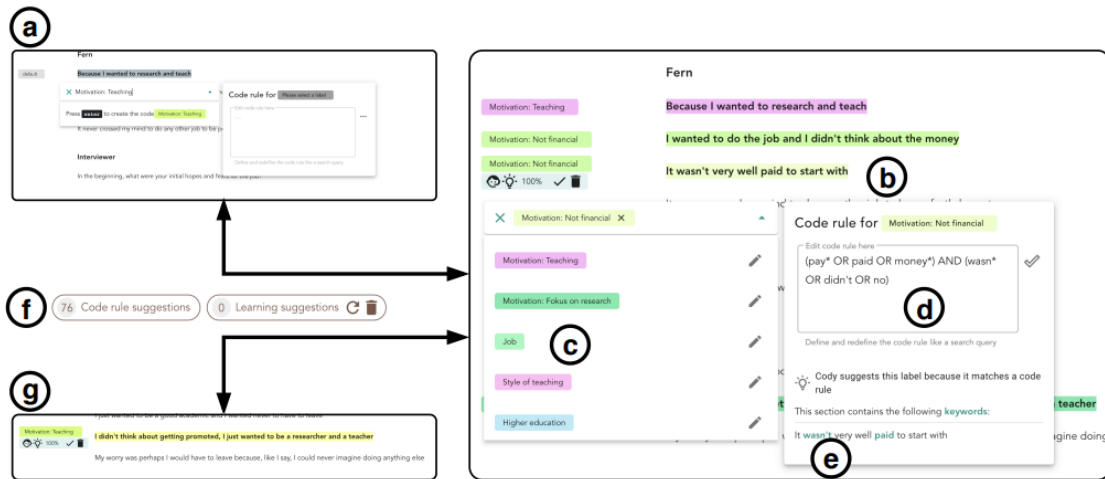


FIGURE 2.2: Rietz and Maedche (2021)'s Cody application

For supervised ML-based code generation, Rietz and Maedche (2021) suggest that (supervised) ML-based approaches are less useful for more ‘open coding’, in which raw data is labelled with newly devised codes. Perhaps this restriction exists because ‘Cody’ cannot predict new codes. Instead, it classifies unseen text segments based on user-inputted codes. Gebreegziabher et al. (2023) add to this saying that supervised ML might not be suitable for QC, because of different objectives (ML models usually used for unambiguous and deductive task domains) and the dynamic nature of coding. Plus, ML-based methods are very likely to have a cold-start problem due to the limited amount of available training data (user-inputted codes), especially at the start of the process. Another flaw indicated by the creators of ‘Cody’ is that imprecise or incorrect pattern rules can cause errors to propagate, resulting in wrong ML code suggestions. Gebreegziabher et al. (2023) also comment that since ‘Cody’ uses a supervised learning model, the code classifier is a ‘black box’, limiting the explainability of the code suggestions and user control.

### Unsupervised: LLM

A different approach to assisting QC with AI is using LLMs. An example of such a system is Gao et al. (2023)'s ‘CollabCoder’, which leverages OpenAI’s ChatGPT. CollabCoder provides AI assistance in several ways. For (independent) open coding, the UX researchers are provided AI-generated code suggestions when adding a code for a particular text segment (see Figure 2.3 for illustration). The ChatGPT code suggestions are generated in two ways: 1) when the user asks for ‘descriptive’ codes, the model is given the currently selected text segment and prompted to

“Create 3 general summaries for [text], within 6 words”

2) when the user asks for ‘relevant’ codes, the model is also given the codes previously inputted by the user and prompted to

“Identify the top 3 codes relevant to this [text] from the following code list:  
[numbered list of user’s code history]”

The AI output is generated using input from the UX researcher only in the second method. Besides inputting codes, the researcher can add supporting evidence for the given label by

selecting relevant keywords or phrases from the raw data cell. Lastly, a certainty level between one and five can be given to the code label. All the codes are added to a “code book”, which is essentially an overview list.

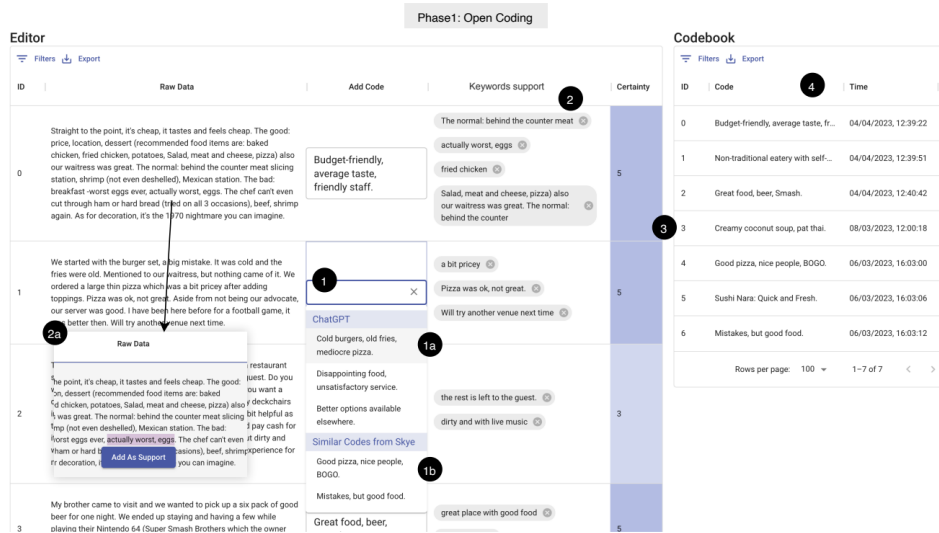


FIGURE 2.3: Gao et al. (2023)’s CollabCoder system

For the suggesting of codes, the AI model is also given additional requirements, such as “6 words or fewer”, “no duplicate words”, “be general”, and “three distinct versions”. After finishing labelling the text segments with codes, CollabCoder allows for the grouping of codes. This can be done manually by the user or the user can request the GPT model to create code groups (see Fig. 2.4). In the latter case, the model is prompted to

“Organise the following codes into 5 thematic groups without altering the original codes, and name each group: [numbered codes]”

Users can request a regeneration of groups or rename and modify the code groups. In this way, researchers are given a starting point for clustering and organising codes and themes. In addition to the prompts, the GPT model is also provided with a desired result format.

Interview-based evaluation by the authors suggests that CollabCoder’s AI suggestions helped with reducing cognitive burden during the independent coding, giving them reference points and helping with data filtering. However many respondents criticised the system for producing too detailed summaries, making the suggestions not directly usable. Additionally for UX research, if the context or other nuances in the user’s behaviour are not explicitly mentioned in the text, the model will not be able to take this into account for its output. Possibly missing out on nuanced or intricate details, producing codes that don’t capture the text segment’s underlying essence.

CollabCoder’s unsupervised approach where the LLM generates code suggestions based on solely the text segment is not limited by user-inputted codes like Gebreegziabher et al. (2023) and Rietz and Maedche (2021). This allows the AI assistance of CollabCoder to be more flexible. Nonetheless, allowing for more freedom in AI-generated output may result in code suggestions that are less representative of the coder’s mental model or ideas. Such concerns are reflected by the study’s respondents expressing a preference to first read the raw data without viewing the GPT suggestions due to fear of it influencing their thinking process. Gao et al. (2023) also warn about the potential reliance on AI assistance,

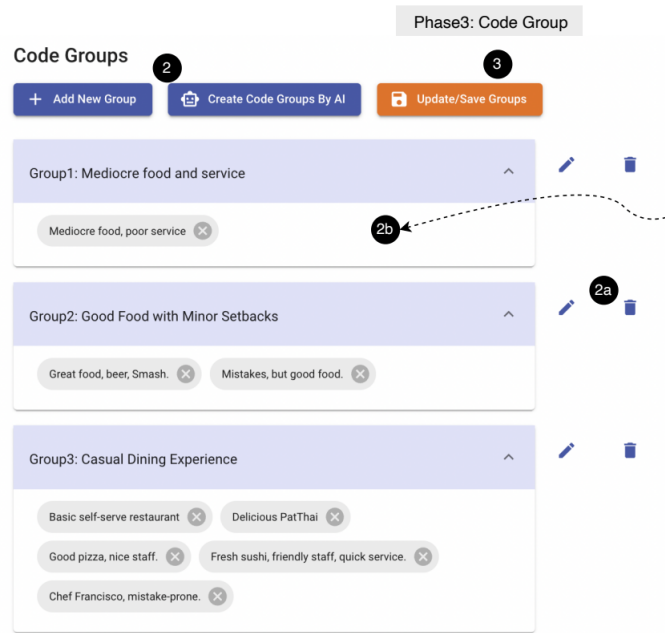


FIGURE 2.4: Generated reports of Gao et al. (2023)’s CollabCoder system: grouping functionality

especially when there’s limited time. E.g. some participants skipped discussions and one participant commented that the AI model seemed to dominate the clustering of codes, dictating the process and approach.

The code suggestions based on text segments do not require input from the UX researcher. However, for other types of AI-generated suggestions, user input is utilised in the form of manually inputted codes. For one type, the input from the UX researcher is handled similarly to rule-based QC assistance, where the researcher’s code history is directly used to code unlabelled segments. The study participants overall found the solely AI-generated code suggestions more useful (deemed more accurate and relevant) than those selected from the user’s own code history. However, they also commented that the latter might be more useful for datasets with less differing contents. More on the incorporation of input from the researcher for the AI output, the authors suggest a limitation of their approach is how it doesn’t incorporate other forms of user input, like user intent in the form of questions. The authors suggest that their approach is limited in that it does not account for other forms of user input from the UX researcher to incorporate into the AI output, such as user intent expressed through questions. Gao et al. (2023) emphasize the importance of human involvement in AI-assisted qualitative coding, to guide the AI model by supplying the nuance, context and deeper understanding that the AI may lack. As well as strategic considerations, such as the level of specificity in code selection.

Lastly, the unsupervised LLM approach abstracts the code from the input text, and without adding highlights it removes the context to some extent. Should the user want to re-examine or evaluate why certain codes were inputted, they would have to re-read the raw data, which can be quite some text if the segment is a whole paragraph. Even though the keywords support can help with this, such support has to be manually added.

### 2.1.3 Summarisation

Another relevant application domain of AI assistance is summarisation. The objective of summarisation is to save time and effort in processing large texts, by condensing them. Condensing textual data can also be seen as a form of data filtering. Moreover, it has similarities with producing summarising overviews of interview highlights for debriefing during the UX testing test day (see Chapter 3).

There are two main types of summarising, namely *extractive* and *abstractive* summarisation. The former refers to extracting relevant key phrases and concatenating them, whereas the latter generates new phrases (Jung et al., 2023). The approach to AI-driven abstractive summarisation is similar to unsupervised QC assistance. The difference between the two AI applications lies in the amount of input and output (few-word labels for QC and multiple phrases for summaries) and the placement of the output in relation to the input, where the summaries are often placed more separately in relation to the input, e.g. below the long input text. Open coding and summarisation are in essence quite comparable tasks, which is nicely illustrated with Gao et al. (2023)'s statement on the difficulties of open coding: "Independent open coding is a very cognitively demanding task, because it requires understanding the text, identifying the main idea, creating a summary based on research questions, and formulating a suitable phrase to convey the summary". Perhaps due to its general application domain, more research has been done on the approaches rather than the creation of specific systems. Hence, in this section, the diverse methods for summarising (supervised vs. unsupervised ML-based) and adding constraints (user inputs, e.g. queries to be answered or keywords to be included) will be discussed rather than specific examples illustrating these approaches, like was done in previous sections. Studies suggest that incorporating user inputs as constraints for the output can improve the performance of unsupervised AI-driven summarisation.

Similarly to AI-augmented QC, there are two main procedures to AI summarisation within the literature, namely supervised vs. unsupervised-based ML. Fine-tuned supervised models trained on large summarisation datasets (with article-summary pairs) and often concerned specific domains. The current leading approach for summarisation is the usage of unsupervised LLMs with prompt-based interfacing, such as the GPT models, which are not trained for particular tasks. Goyal et al. (2023) conducted a study comparing these two approaches and demonstrated that all respondents had a preference for the GPT-produced abstractive summaries. Furthermore, the 'old' paradigm necessitates considerable training before it's applicable for new domains (other domains than that of the original training data).

Different approaches to constrain or control AI-generated summarisation have been investigated, which can often be compared to different forms and extents of user input. Examples include *keyword-based* (often specific entities or events directly mentioned in the input data), *aspect-based* (more high-level topics that can be common across datasets) and *query-focused* (Goyal et al., 2023). Unconstrained is called *generic summarisation*. Results of Goyal et al. (2023)'s study demonstrate that the GPT model performed better than the fine-tuned model for keyword-focused summarising. However, both models produced poor results for aspect-based summarisation.

Jung et al. (2023) did more research on administering constraints to the unsupervised ML-based summarisation systems and doing so with the user via an interactive interface. The article utilised keyword- and *key phrase-based* constraints, namely word positives and part positives. Word positives are imperative words that should be included in the summary, and part positives refer to topic phrases whose formulation may change, as long as the in-

formation is maintained. The system uses an AI-based PoS tagger to identify and highlight (in raw data) potential keywords, which the user can check, edit (e.g. delete highlight if the keyword is not relevant), or add new highlighted keywords. Interfaces for highlighting keywords and keyphrases are separate. After completing this constraint highlighting task, users can press the ‘generate’ summary button. The whole process can be repeated till satisfaction. Evaluation results indicate that constraint-sensitive models can increase the performance of abstractive summarisation systems. In addition, respondents found the interactive interface to be more helpful than the standard generative AI summarisation (without constraint tagging).

Zhang et al. (2023) investigated the usage of ChatGPT for extractive summarisation. The paper concluded that the LLM still underperformed compared to the supervised models in terms of ROUGE scores (comparing produced summary with golden standard summary, per n-grams). Nevertheless, the authors found that first conducting extractive summarisation and then generating an abstractive summary of that (instead of only abstractive summarisation), improved ChatGPT’s adherence to the original data.

Despite the recent advances and tactics to tackle models’ limitations, current automated summarisation systems are usually unsatisfactory and inadequately summarise the input text. This is because the summaries are still susceptible to errors and have a tendency or potential for hallucination, providing incorrect output that doesn’t correspond with the input (Fok et al., 2023). Plus, summaries don’t allow users to quickly interact or review the original data, thus providing limited context of the given output.

#### 2.1.4 Skimming

Another relevant application domain of AI is skimming assistance, where AI is used to filter the textual data to be processed, in order to save time and effort. Skimming is a faster form of reading or textual data processing. It is a cognitively demanding task where readers swiftly review a paper’s content, to get a general overview and understanding of it. While skimming, one usually focuses on information relevant to one’s objective, thus filtering the data. To do so effectively, one has to make strategic choices of what to read, where and when to stop reading. During the process, readers continuously build (and adjust) a mental model of the paper’s relevant content, integrating information across sentences and sections read.

Little research has been done on assisting skimming with AI, but one such system employs supervised ML, which will be discussed in this section.

Fok et al. (2023) developed Scim, an application specifically intended to help with skimming scientific papers. Scim attempts to help users filter data by redirecting the user’s attention. The application does this by highlighting the potentially most relevant sections. Scientific papers all follow a certain structure with content related to e.g. objective, method, result, etc. Scim first segments the text into sentences then classifies the segments according to the earlier mentioned structure elements and highlights them with different colours. Thus, highlights are given in context (in the raw data) as well as in an overview on the side (see Figure 2.5). In addition, a classification probability score is shown for each highlighted segment. Supervised ML classification was employed, hence extensive training was required, using a dataset of scientific papers with manually-curated ground truth labels. The labels were further fine-tuned using i.e. keyword matching with specific words (e.g. we, our, this paper, and their aliases for detecting intent and objective). This fine-tuning is similar to the keyword constraints used for summarisation, however, the input is given by the re-



searcher (when training the model) rather than the user. Such pre-determined constraints are perhaps applicable to the task of skimming since scientific writing has a distinct style and structure. However, in the case that the fixed keywords have a different meaning in the context, then it could be that this is lost due to the researcher’s initial interpretation. The authors comment that this approach (in contrast to summarisation) allows the readers to utilise the traditional visual and structural landmarks of the paper (e.g. headers), naturally retaining the context.

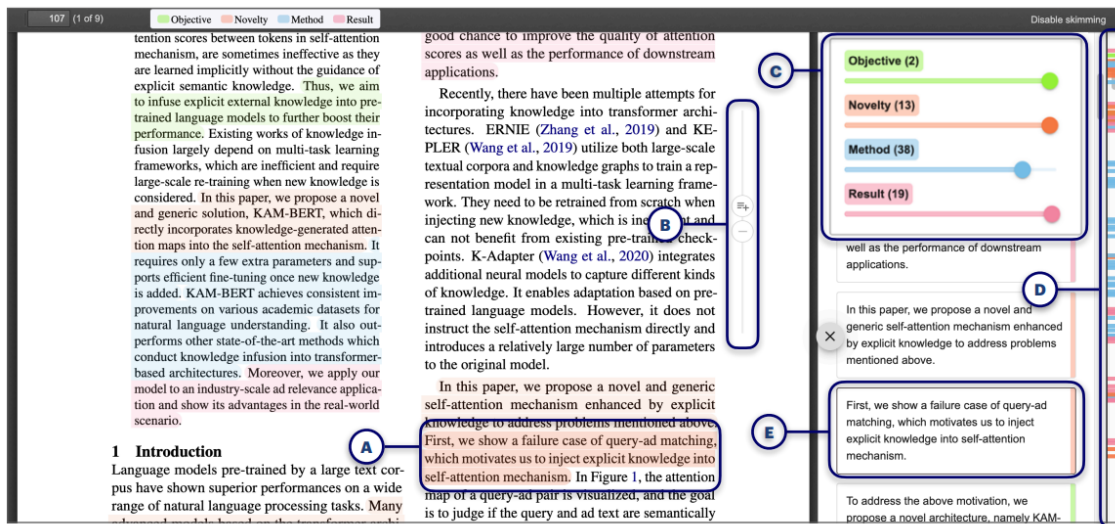


FIGURE 2.5: Fok et al. (2023)’s Scim system

### 2.1.5 Takeaways for AI textual data filtering assistance

There are various ways in which AI assistance can be applied to help filter textual data, to save time and effort. The discussed application domains are qualitative coding (QC), summarisation and skimming.

Within the QC domain, we described two systems that make use of NLP-rules, where one application combines it with supervised ML. For rule-based code suggestions, the AI assistance is restricted by the modelled syntax. Supervised ML approaches have cold-start problems and are generally more suited for unambiguous application domains. The AI-generated codes are presented in the physical vicinity of the corresponding text segment, retaining its context to some extent. We also discussed the unsupervised LLM-based approach of CollabCoder. Although CollabCoder does not require manual labelling of raw data and is not limited by pre-determined rules, the freedom can result in code suggestions that are less representative of the user’s mental model. This is accompanied by concerns about possible harmful AI influence and reliance on AI assistance.

For summarisation, the literature contains research on both unsupervised and supervised approaches. Supervised summarisation requires lots of training and manual creation of golden-standard summaries, and is often domain-specific. Unsupervised approaches don’t necessarily require lots of training and are more generalisable, but have a higher tendency to hallucinate. Both approaches are said to still be susceptible to errors, resulting in incorrect outputs. Plus they provide limited context, as the AI output is generally displayed spatially separate from the original input text. It is advised to use user input to constrain and control

the AI-generated summarisation.

AI-driven skimming assistance employs supervised ML learning for classification, thus requiring training and manual labelling. On the other hand, how this assistance is given naturally retains context.

## 2.2 Cognitive Load

In this section we will discuss what the construct of cognitive load (CL) entails, the three types of cognitive load, factors affecting CL, and lastly we'll further elaborate on how to quantitatively measure CL.

A big portion of the literature on CL is written in the context of learning and education. The content of this section has been slightly reformulated to place the presented information in the context of UX testing. For example, instead of talking about learning tasks, we merely say tasks or instead of instructional design, we refer to the design of materials and systems.

### 2.2.1 What is cognitive load?

Paas et al. (2004) define *cognitive load* as “mental activity realized simultaneously with working memory”. CL is generally considered a multidimensional construct that represents the load that performing a particular task imposes on the cognitive system of a learner (an individual)” (Meshkati, 1988; Paas & Van Merriënboer, 1994; Paas et al., 2003; Yeh & Wickens, 1988).

Paas and Van Merriënboer (1994)'s model presents CL to be composed of so-called ‘causal-’ and ‘assessment factors’, which are elements that influence and are influenced by cognitive load, respectively. The causal factors are task and subject characteristics. Task characteristics that have been identified in the research are the amount of information, information and task complexity, task novelty, task structure, time pressure, and pacing of instruction (information). Novel tasks performed under high time pressure are associated with high CL. Subject characteristics encompass expertise level and experience (training), cognitive capabilities, psychomotor skills and physical abilities, age, and arousal state (Meshkati, 1988; Paas & Van Merriënboer, 1994; Paas et al., 2003). The assessment factors comprise mental load, mental effort and performance, as well as controlled- and automatic processing. *Mental load* is said to indicate the expected cognitive capacity demand of a task, taking a subject's current knowledge of the task and its characteristics into account.

### 2.2.2 Types of cognitive load

In literature, a distinction between three types of cognitive load (on working memory) has been made (Mutlu-Bayraktar et al., 2019; Paas & Sweller, 2014; Sweller et al., 1998). According to the cognitive load theory (CLT), these are intrinsic cognitive load (ICL), extraneous CL (ECL) and germane CL (GCL). ICL concerns the *complexity* of a task and one's prior knowledge, ECL concerns the increase in mental processing and effort due to inappropriate design of used materials and systems, and GCL occurs during the formation and regulation of mental structures or the effort expended that contributes to knowledge construction (i.e. making links between concepts, and thinking of themes). Within CLT, ECL is often labelled as unnecessary and undesired. Hence, it is argued that ECL should be limited as much as possible to free capacity for more learning-related processing, considering people's limited working memory capacity (C.-Y. Chen & Yen, 2021; Sweller,

2010; Sweller et al., 2011). Learning-related processing refers to GCL, which should be maximised. Thus, together –ICL, ECL and GCL– form the total cognitive load that could occur.

### 2.2.3 Factors for cognitive load

The literature on CL mentions various factors affecting an individual’s extraneous cognitive load during performing tasks and related theories. We focus on ECL as that will depend on the AI assistance system that will be implemented. These include the segmenting principle, seductive details effect, split-attention effect, coherence principle, spatial contiguity principle and temporal contiguity principle (Beege et al., 2017; Jan et al., 2016; Makransky et al., 2019; Mutlu-Bayraktar et al., 2019; Rop et al., 2018). What these factors entail is described below:

- *Segmenting principle*: It’s said that people learn better when information is presented to them in segments, instead of a big continuous stream. This is related to how the amount of information that’s given as input into working memory (at a point in time) can be controlled. More input provided at the same time means a higher CL because more information has to be processed simultaneously in working memory.
- *Seductive details effect*: This theory describes that irrelevant information decreases people’s comprehension. The more irrelevant information is included, the higher the CL.
- *Split-attention effect/spatial contiguity principle*: This principle states that related pieces of information can be better processed when presented spatially together rather than separated unless one of the pieces of information is unnecessary.
- *Coherence principle*: It’s said that the provided information should be consistent and align with established learning objectives. If the information is not consistent, the CL will be higher.
- *Temporal contiguity principle*: This principle describes that related pieces of information should be integrated and presented in a synchronised manner. If the related information is not given at the same time, the CL will be higher.

Moreover, Rop et al. (2018) state that time also plays a role in the CL when performing a task. In a limited amount of time, elements need to be processed faster. As a result, a normally cognitively undemanding task can become very demanding when time is limited (i.e. working under time pressure).

Lastly, Örün and Akbulut (2019) and Salvucci et al. (2009) mention that (concurrent) multitasking increases mental effort and thus cognitive load as goals of different tasks “compete” at the same time to control cognition and to take up the capacity of one’s working memory.

### 2.2.4 Measuring cognitive load

Various methods have been used in previous research to measure cognitive load, which can be done both objectively and subjectively (Mutlu-Bayraktar et al., 2019). Examples of objective measurements are physiological measures, such as heart rate, pupil dilation, and brain scans, like fMRI, EEG, etc. Additionally, performance-related metrics, i.e. time-on-task and task performance have also been utilised.

Subjective measurements are generally conducted via self-evaluation (i.e. self-reported mental effort expenditure) using questionnaires. Various questionnaires exist, but some that are well-known and reviewed are the Paas scale (Paas & Van Merriënboer, 1994), the NASA-task load index (Hart & Staveland, 1988) and the Leppink questionnaire (Leppink et al., 2013). They measure CL as follows:

- Paas: 1 question for rating mental effort expended to perform a specific task on a scale of 1-7; doesn't differentiate between different types of CL.
- NASA: 6 dimensions/items to be rated on an 18-point Likert scale. Dimensions consist of mental demand, physical demand, temporal demand, performance effort and frustration level
- Leppink: 10 items to be rated on a Likert scale of 0-10, composed of 3 ICL items, 3 ECL items and 4 GCL items

Subjective measurements are the most commonly used methods, although Brunken et al. (2003) state that objective measurements are more valid and reliable to measure CL.

### Chapter conclusion

In this chapter, we have examined related literature on the fields of AI for textual data filtering assistance and cognitive load.

Firstly, we have discussed several relevant AI-based techniques and approaches for textual data filtering assistance that aim to save users time and effort. Each approach has its benefits and limitations; e.g. supervised and NLP rule-based learning is generally more rigid in approach compared to unsupervised learning and is more time-consuming to train, although unsupervised learning can produce too unconstrained outputs and thus has a higher risk for hallucination and incorrect outputs. Above all, unsupervised LLMs appear to be highly suited for processing textual data. Furthermore, their flexibility, little requirement for intensive training and manual labelling make them more appropriate for the flexible procedures and desired output of the UX testing workflow than the more rigid supervised methods.

For AI-assisted textual data filtering, user input can be used to constrain and control the AI output, or to help capture factors and nuances of complex human behaviour that may be challenging for AI to model. Furthermore, the extent to which the AI output is given context or is placed in context can have an impact on the (re-)evaluation of AI output.

Next, cognitive load (CL), if simply explained, is the mental activity or effort expended when performing a particular task. Cognitive load theory describes three types of CL –intrinsic, extraneous and germane CL– that together use up people's limited working memory capacity. Various factors affecting ECL, which depends on the AI assistance system that will be implemented, have been determined within the literature. These factors are described by the segmenting principle, seductive details effect, split-attention effect, coherence principle, spatial contiguity principle and temporal contiguity principle. Limited time and multi-tasking also contribute to cognitive load. Lastly, various ways to measure CL have been established, such as the Paas scale, the NASA-task load index and the Leppink questionnaire.

In Chapter 4 we use these findings to make decisions regarding the experiments we conduct to answer our research question. However, first, we examine the thesis company's UX testing workflow to expand our necessary background knowledge.

## Chapter 3

# UX testing workflow

To be able to answer our research questions and investigate the impact of AI textual data filtering assistance on the UX testing workflow, we need to get a clear picture of said workflow. For the method of UX testing, the standard procedure of the thesis company is taken as the basis. The information was collected through interviews with five UX researchers, who have different amounts of experience in the field of UX (ranging from 2 to 18 years)<sup>1</sup>. A noteworthy characteristic of their work procedures is that most of the tasks are done in pairs, which allows for sharing the workload and having multiple perspectives, which can reduce possible bias through discussion.

Their UX testing workflow is standardised and consists of roughly the following main phases (illustrated in Fig. 3.1):

1. Meeting with the client
2. Preparation
3. Test day
4. Report-making day

For the first phase, ‘Meeting with the client’, the aim is to understand the client’s objectives (client questions), context, and specific requirements for the UX test day. Additionally, the test material—ranging from existing sites or apps to prototypes—is reviewed, and its scope is discussed concerning the client’s questions (e.g. most important use cases).

Next, the UX researchers prepare for the test day and familiarise themselves further with the project objectives, context and test materials by working on the test script, which contains scenarios or tasks and questions for the respondents. The test script functions as a starting point or guideline for the interviews on the test day (especially regarding the interview structure or flow). During this phase, the respondents for the test day are recruited, but this is mostly done by another company and thus excluded from this analysis.

The following phase is the ‘Test day’, during which UX researchers aim to collect the data to answer the client questions. It was determined that the *obslogging* (logging observations) and *debriefing* (discussing findings) tasks during this phase have the highest cognitive load, and are deemed to be the tasks where the most improvement can be gained in terms of efficiency and reduction of workload by using AI assistance. Since the AI data filtering

---

<sup>1</sup>Liem, S.Y. (2023). *Research topics report: The exploration of AI for the UX testing work*

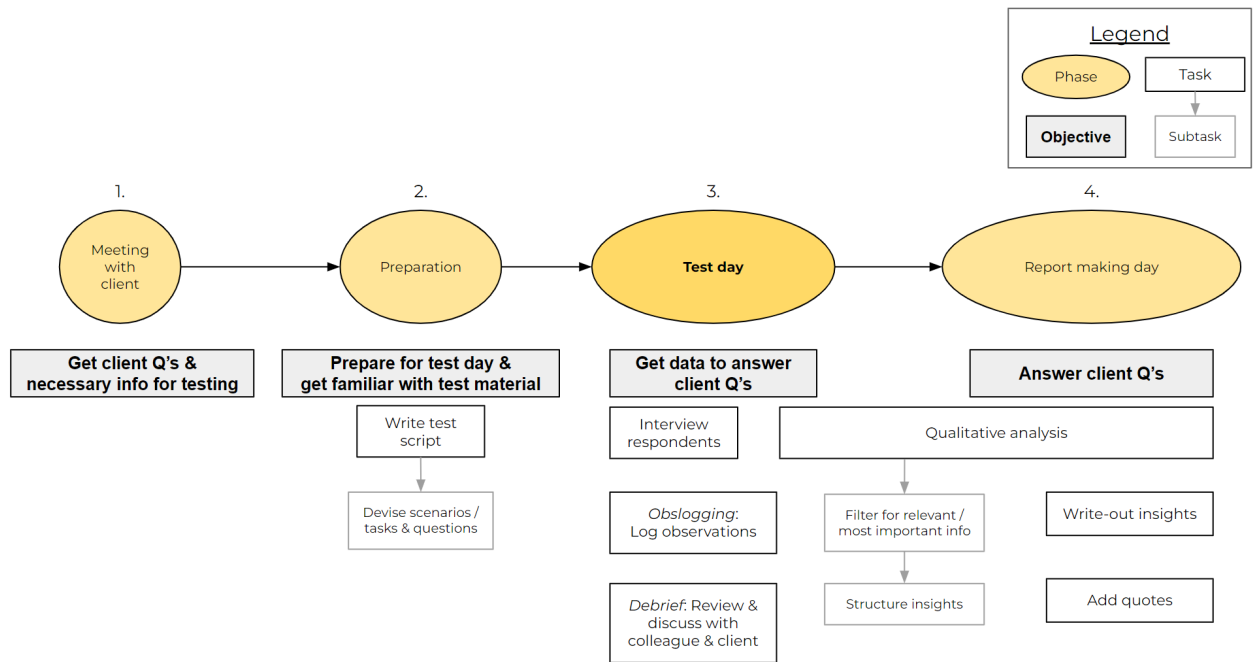


FIGURE 3.1: Visualisation of the UX testing workflow's main phases

assistance will specifically be applied during the 'Test day' phase, the focus of this research will be on this step. Consequently, this step will be further elaborated in this chapter.

The last main phase is the 'Report making day', during which researchers aim to answer the client questions by creating a slide presentation (the report). This phase will also be described in more detail in the coming chapter as the AI assistance used on the test day will have a direct impact on this phase. This is because the data collection step ('Test day') determines the data collected to answer the client questions on the 'Report making day'. Moreover, the qualitative analysis was determined to already start on the 'Test day' and hence overlaps in the last two phases. The UX researchers present the report made on another day.

Although the thesis company's UX testing workflow is standardised, the procedures are somewhat flexible, allowing individual researchers to change how to perform a specific task or use (specific aspects of) a tool or document.

### 3.1 Test day

The objective of the test day is to obtain all the information necessary to be able to answer the client questions. The testing is conducted by two UX researchers, who evenly share the role of interviewer and observer (who does the obslogging) for the interviews with the respondents. Besides the researchers, the people of the client's team are also present, observe the interviews and may take notes (in an online whiteboard prepared by the researchers or some other way). In between interviews, the researchers and the client briefly discuss some highlights or remarkable observations. At the end of the test day, a more extended discussion is held, the debrief.

For the obslogging and debriefing tasks, the researchers have to multi-task a lot; they need

to process and filter data (what they hear and see during the test day), and attempt to get an overview of the key content (the highlights). These underlying processes are often challenging and are experienced to have a high cognitive load.

### 3.1.1 Obslogging

For logging observations or *obslogging*, the UX researchers aim to record relevant information (that is said or seen) that answers the client questions. Additionally, they pay attention to information concerning the user experience of the test material. When observing, the researchers strive to listen attentively, process and filter what is being heard and seen for relevant information, write down that information, label this information, and already process and filter for most important information, note down such highlights, optional recommendations, insight links, other ideas, etc. Some might even already try to organise the most important information. As described, a lot of processes are done simultaneously (multi-tasking), resulting in a high cognitive load. A high cognitive load or a cognitive overload can cause people to struggle to process information or make appropriate decisions (Örün & Akbulut, 2019; Rop et al., 2018; Salvucci et al., 2009). Plus it can be mentally taxing, causing mental fatigue and impairing cognitive performance. One’s typing ability and the talking speed of the respondent can influence the speed and thus the challenge of multitasking.

	A	B	D	E	F	G	H	I	J
1	Klok-tijd (automatisch)	Respondent	Taak	Context	Label	Impact	Observatie of citaat	Video	Inzicht
2							Zit je bij [redacted] op lokatie? Typ in deze kolom H een observatie "start" zodra de interviewer in OBS op 'Start recording' klikt. Dan loopt de tijd-notitie in kolom C (staat ingeklapt) gelijk met de looptijd van de opname. Zo vind je later een quote makkelijk terug in het video-bestand. Een voorbeeld:		
3									
4									
5		R1, Frits, 28 (Desktop)	pre-test	achtergrond			start		<input type="checkbox"/>
6		R1, Frits, 28 (Desktop)	I. Filteren	product-detail pagina			"Oh wat leuk, dit is precies wat ik nodig had. Nou zeg, dat lost echt iets op."		<input checked="" type="checkbox"/>

FIGURE 3.2: Obslog: Template used for logging observations

At the thesis company, a semi-automated Google Sheets template is used for logging observations, called the “Obslog” (see Fig. 3.2). This task is commonly called *obslogging* under colleagues. Column I, with ‘Observation or quote’ (in Dutch: ‘Observatie of citaat’) as the header, automatically changes the font style of a quote indicated by using quotation marks. The time stamps are added automatically, and columns D-G, titled ‘Respondent’, ‘Task’ (‘Taak’), ‘Context’ and ‘Label’, automatically fill in the value of the most recently filled-in row for a new row one is typing in. The three columns ‘Task’, ‘Context’ and ‘Label’ are intended to give observers the space to give context about their corresponding ‘Observation or quote’ cell, where the context gives more information on what the ‘Observation or quote’ is about or refers to. The three columns allow three levels of granularity, however, the exact specificity of each column can differ per project, per researcher and even over time for one observer (see Figure 3.3, 3.4 for illustration). Nonetheless, the ‘Task’ column is intended to have the lowest granularity and ‘Label’ the highest.

Figure 3.3 illustrates how different UX researchers label the same scenario in various ways. For example, the term ‘voorbereiding (...)’ is placed in three different columns by three observers. Moreover, in Fig. 3.3c the observer has not even used the ‘Label’ column at this point. Some researchers sometimes decide to remove this column when they deem it





- ‘!’ for a neutral impact but very interesting information
- ‘?’ for any confusing moments

The UX researchers expressed that the ‘Impact’ column values are selected based on feeling, which also explains how some might use the values differently or with different nuances.

This labelling process is comparable with qualitative coding (QC), where the ‘Task’, ‘Context’, ‘Label’ and ‘Impact’ columns can be seen as labels or codes for the observation or quote. The labels used for obslogging and QC have similar characteristics, such as having different levels of information, depending on the coder and their needs, and their iterative nature as they emerge and evolve over time. Also, labelling during the obslog task and QC are both forms of textual data filtering. However, iterative labelling is not done as thoroughly during observation due to the difficulties of multi-tasking and time constraints. The challenges of open coding, specifically how cognitively demanding it can be, are also applicable to the thesis company’s method of observation logging.

The columns ‘Video’ and ‘Insight’ (‘Inzicht’) are occasionally used to highlight a valuable moment or quote or to note down any interesting thoughts on recommendations, insight links, reasoning or even highlights that are useful for the debriefing and later analysis. Quotes are an account of what the respondent has said, which the researcher thinks may be useful as support for insights in the report. Observations describe what the respondent is doing or rephrase what has been said. The amount and specificity of the observations may change over time; e.g. as the interviews progress, researchers get a better grasp of what’s relevant or not and how to describe best what an observation is about (labelling). The extent of adding extra information such as labels depends on their workstyle and preference, typing speed and possibly even multi-tasking ability. If a researcher is a slow typer or has more difficulty multi-tasking, fewer labels might be added, affecting the ease of later analysis.

Depending on the ease of obslogging and time available, researchers even try to note down highlights in the “Issue list” (see Figure 3.5). The issue list is a separate tab of the Google Sheets document; a space separate from the ‘obslog’ sheet for an overview of the most relevant points or issues mentioned. The issue list is aimed to make the debriefing and later analysis easier, saving time and effort (during the next steps). However, most researchers are not able to do this during the interview or in between interviews due to lack of time.

	A	B	C	D	E	F
1	<b>De titel van het project</b>					
2						"Met welke doelen van de organisatie en/of jullie team helpt dit jullie verder?"
3	<b>Impact</b>	<b>Wie schrijft?</b>	<b>Geschreven?</b>	<b>Nagelezen?</b>	<b>Inzicht titel</b>	<b>Relevant doel van team / organisatie</b>
4						
5						
6						

FIGURE 3.5: Issue list template

A benefit of active observation is that it can help keep the researcher (more) focused on the interview, as one has to actively listen and process the information to write it down in a sensible manner. Compared to automatic transcription of the audio, the researcher already filters out the relevant (key) content and makes coherent phrases from respondents’ inco-

herent half sentences. This makes a re-visitation of the data more efficient. Additionally, researchers can also take into account what they see, which a transcription cannot.

The mentioned challenges, potential areas for AI assistance and other remarks relevant to the potential of AI for improving this task of the UX testing process are summarised below:

- Underlying processes: data processing and filtering for relevant (key) information, some form of qualitative coding, and sometimes already summarising the relevant information (extract highlights)
- Challenge: high cognitive load due to multi-tasking, meaning a risk for mental fatigue and (cognitive) task performance impairment
- Potential of AI: decrease cognitive load, also saving time and effort; assist observer by supporting any of the data filtering tasks (decreasing amount of multi-tasking), such as labelling
- Remark: active note taking can help the observer stay focused, and the researcher can generally “automatically” filter and process what’s being said AND seen for relevant information.

### 3.1.2 Debriefing

During the debrief, the researchers aim to discuss the most important observations and insights with the client and colleagues. Before the start of the discussion, the researcher tries to remember or remind themselves of the highlights, and thus what to discuss. In other words, they try to get an overview of the key content, which is akin to summarising the content. Depending on the amount of time before the debrief, the researcher’s work style and the usage of the issue list during the interviews, the extent of the preparation and strategies used may differ. UX researchers may write down highlights from the top of their mind, whilst scrolling through the obslog, or not.

During the debriefing, researchers actively listen, react, add insights, and guide the conversation to examine important topics and insights. One researcher leads the conversation whilst the other records the discussed information in the issue list. When the information is written down, the researcher already structures it to some extent if they have the time and energy, e.g. per topic, insight, flow step, theme, etc. Although the columns in the issue list are named and thus give some template for structure, it is generally ignored and the sheet is freely used. Instead, the researcher may organise the data as they see fit, using bold text for headers and such. If anything was already written in the issue list, researchers write down the debrief notes separately, under a debrief header. If the client actively wrote down useful observations or the researcher finds it more suitable or comfortable, the debrief notes may also be recorded on the online client whiteboard.

The biggest challenge for the discussions lies in remembering the most important points to discuss. The ease of this task relies on having an overview of the key content, either in mind or on record. This depends on memory and how meticulous and/or structured the observations were documented. The cognitive load experienced during obslogging can also affect this, as a high load can lead to mental exhaustion and difficulties in thoroughly recording, processing, and organising observations. Memory might also be affected by this; several researchers have experienced a blackout when recalling the highlights during the debrief.

Researchers have different approaches to acquiring an overview of key content: some depend mostly on their memory of the clients, colleagues, and themselves, assuming that what they remember is the most crucial information. This approach carries the risk of generating an incomplete or skewed view. People can get lost in highlights or details that stay top of mind. If the researchers collected (and organised) highlights in the issue list, this could be used. However as mentioned earlier, the usefulness of the issue list may differ. Lastly, some researchers skim or scan the obslog, filtering on the impact or other labels. The ease of doing so again depends on how thorough the observer was.

The mentioned challenges, potential areas for AI assistance and other remarks relevant to the potential of AI for improving this task of the UX testing process are summarised below:

- Underlying processes: data filtering for relevant key content, and creating an overview of this
- Challenge: finding the key content to discuss can be very challenging and has time constraints
- Potential of AI: make the task easier and save effort; assist observer with data filtering, like labelling or highlighting the (relevant) information, creating an overview of highlights
- Remark: how challenging the debrief and check-ins are depends to quite some extent on how the previous task of the observer went, during which cognitive overload or a high load can occur

## 3.2 Report-making day

The final objective of a UX test project is to create a report that conveys the most important and relevant insights in a structured way so that the resulting story is comprehensible, compelling, persuasive and digestible. To achieve this, more qualitative analysis has to be done, finding patterns and themes within the observations and insights, and the connections between them. On the report-making day, the researchers recall the key points to discuss and include in the report, discuss those highlights with their partner, record or visualise the highlights, related thoughts, insights and links, and try to structure the data. These steps are akin to the qualitative analysis undertaken during the test day. In addition, the insights, themes and patterns found are written out and formulated in a coherent way. To support their findings, researchers add quotes. Recommendations corresponding with the insights are given, frequently with existing examples. The analysis and report-making is an iterative process, in which the researcher goes back and forth between all the steps. It's generally during this process, especially for complex datasets, that researchers discover the appropriate themes and structure for their insights. This helps them to effectively communicate their findings and provide a clear takeaway message or answer to any client questions.

### 3.2.1 Qualitative analysis

Similar to the memory-based extraction of the key content during the test day, the UX researchers start with a sort of mind dump (e.g. in the issue list, the online client whiteboard, or physical sticky notes); in other words, they use a top-down approach for the qualitative analysis. Unlike when using a bottom-up approach, the researchers utilise their

‘human filter’ to reduce the amount of data to go through to find the most important points, thus saving a lot of time and effort. Even though the recalling of highlights is similar for the debriefing, during the report-making day the researchers have the time to revisit the data to fill the gaps. They do so using the obslog, and possibly the issue list and the online client whiteboard. Focusing on the obslog, the UX professionals use the labels (‘Respondent’, ‘Task’, ‘Context’, ‘Label’, ‘Impact’) to scroll through the data in a more focused manner. In addition, the Google Sheet’s ‘Find’ functionality is heavily used, searching for specific keywords or synonyms they remember. To employ the ‘Find’ function, they first have to have a rough idea of what they’re looking for, and then also be able to formulate the correct wording. Researchers often use search words related to the ‘Task’/‘Context’/‘Label’ columns, somewhat remembering that something interesting or useful occurred during that task or action. Despite having tools to filter and reduce the obslog data, the researchers may still “get lost” in the sea of data. This is especially the case if the labels are insufficient or if their memory or mental model of the insights and themes is a bit muddy.

Some researchers may start the QA by going through the issue list, but that is only possible if it was filled well during the test day.

How the UX professionals note down or visualise the insights during the analysis can differ per colleague’s workstyle preference and the observation-insight data. The researchers may work in the issue list of the obslog document, where the majority of the collected data is located. Otherwise, if they feel that the linear and rigid format of a Google Sheets document is not suitable, they might switch or start mapping insights, overarching themes and relations between them in a more flexible and visual-based online whiteboard. Nonetheless, using the issue list space is the most common method.

When structuring the information written down, the researchers often cluster the observations or insights by i.e. similar (sub-)topics, such as findability or layout. Structuring also already happens during the initial stages of analysis / the mind dump. Although the initial organisation is likely to change to a certain extent since qualitative analysis is an iterative process, it already gives a starting point for writing and further analysis.

The mentioned challenges, potential areas for AI assistance and other remarks relevant to the potential of AI for improving this task of the UX testing process are summarised below:

- Underlying processes: data processing and filtering for relevant key content, and structuring content
- Challenge: finding the key insights and structuring the insights in a logical and comprehensible manner can be very challenging and time-consuming
- Potential of AI: make the task easier and save time and effort; similar potentials for debriefing, and assist with clustering of observations/insights
- Remark: unlike during the test day, researchers have (more) time to fill in potential gaps in their overview of key content

### **3.2.2 Report making**

The insights are written down and presented in a Google Slides presentation. Finding quotes to add as support to insights, researchers apply a similar method of focused searching the obslog during the analysis. Sometimes researchers may search per respondent for useful quotes, for a more even quote distribution or to check what other respondents said about a

certain topic. Additionally, more explorative scrolling is done when one has more difficulty finding a nice quote that summarises or describes what one wants to say.

The mentioned challenges, potential areas for AI assistance and other remarks relevant to the potential of AI for improving this task of the UX testing process are summarised below:

- Underlying processes: filtering for relevant key content (quotes) and text formulation
- Challenge: finding suitable quotes that best support the mentioned insights and are evenly distributed over respondents
- Potential of AI: make the task easier and save time and effort; data filtering in terms of finding quotes
- Remark: unlike during the test day, researchers have (more) time to scroll through the obslog in search of useful quotes. Dependent on how well the researchers noted down relevant quotes by respondents

For the whole workflow, the UX researchers may feel time pressure as they try to complete the tasks in the assigned hours even if the project and the obtained insights turn out to be more complex or challenging to analyse than expected.

### Chapter conclusion

The crucial underlying process during the ‘Test day’ and ‘Report-making day’ of the UX testing workflow is data filtering. Filtering the collected data for relevant content can be challenging, more so due to the project-based time constraints UX researchers have. Therefore the UX testing workflow could be improved (in terms of time, effort and cognitive load) by utilising AI to aid with data filtering on the ‘Test day’, which then also impacts the data filtering on the ‘Report-making day’. Providing AI assistance for the tasks of *obslogging* and *debriefing* is likely to lead to work-process improvement on the ‘Report-making day’ as well.

For *obslogging*, the multi-tasking and thinking-of/formulating the ‘Task’, ‘Context’ and ‘Label’ labels are especially challenging and taxing. Hence, helping the observer with labelling the ‘Observation or quote’ cells is a good starting point to examine the potential of AI for workflow optimization. For *debriefing*, the challenge of data filtering is expressed in the form of finding (an overview of) the key content to discuss. Utilising AI to provide an overview of the highlights (key content) of the interviews could help UX researchers. This way, AI can be applied to help with faster (more efficient) filtering, possibly reducing the cognitive load and difficulties of the tasks and potentially improving the UX testing workflow.

The information on the UX testing workflow provided in this chapter, combined with the related literature on AI and cognitive load, was used to make decisions regarding e.g. the experimental set-up and the design of the AI assistance. Such design choices are presented in the following chapter.

# Chapter 4

## Methodology

The goal of this research is to explore how AI textual data filtering assistance influences the tasks of noting down observations (obslogging) and debriefing for UX researchers (RQ1). Additionally, the objective of this study is to understand how UX researchers experience AI textual data filtering assistance during a UX testing test day and why (RQ2). Lastly, the study aims to assess the extent to which AI textual data filtering assistance can alleviate cognitive load during the UX test analysis process (RQ3). To achieve these goals, experiments are conducted during which we simulated a UX test day, more specifically the tasks of noting down observations (obslogging) whilst watching interviews and holding discussions on the interviews afterwards (debriefing).

The participants are subjected to two test conditions: executing the tasks of obslogging and debriefing with (A) and without AI assistance (B). Next, we observe how they perform the tasks (RQ1) and inquire the participants about their experience and thoughts regarding the AI assistance (in relation to performing the tasks and compared to without AI assistance; RQ2) during the post-test interview. Moreover, we measure the respondents' cognitive load during the different tasks and set-ups (RQ3).

### 4.1 Participants

The participant pool consisted of eight employees of the company, who have experience conducting UX testing test days. Participants include UX researchers with varying amounts of experience: an intern, a junior, two mediors and two seniors. As well as two UX designers, namely a junior and an intern. Interns were included to diversify and expand the participant pool, representing novice researchers. Similarly, UX designers (with UX testing experience) were also allowed to participate. The participants' age ranges from 22 to 56. Plus, the male/female distribution was 50:50. UX researchers who have produced the project data used to create the AI assistance are excluded, as well as the company supervisor, since they have too much knowledge of the experiment. The participants were recruited by sending them a Slack message and a Google Calendar invite, along with an informed consent form.

A within-subject design is employed due to the limited number of participants and to restrict the effect of individual characteristics on the results.

## 4.2 Variable Manipulation & Measurement

For this research, the independent variable is the presence of AI textual data filtering assistance. For the first and second RQs, the UX researchers' behaviour and experiences can be seen as the dependent variables, respectively. The dependent variable for the third RQ is cognitive load.

### 4.2.1 Independent variable: AI textual data filtering assistance

For this experiment we settled on binary factor levels for the presence of AI textual data filtering assistance, resulting in two test conditions, namely performing the tasks of obslogging and debriefing (A) with and (B) without a form of AI textual data filtering assistance. For the without condition, the participant utilises the standard Obslog document they are familiar with (see Chapter 3). For the with condition, the Obslog document is slightly altered to make room for the AI assistance.

It was decided to integrate the AI assistance into the companies' standard tools to adhere to the familiar work procedures. The idea behind this is to make it easier to try out the AI assistance without having to put much effort into learning new tools. The AI assistance is produced utilising the OpenAI API. The AI assistance was formed based on and/or inspired by related works, the needs expressed in the background section and the feasibility of implementation. The employment of generative AI and LLMs (OpenAI API) was decided for the ease of implementation for both tasks, considering the limited time available and the focus on exploring the impact of AI assistance rather than developing and testing an ideal AI technology or tool. In addition, the flexibility and high suitability of LLMs for textual data was preferable to the more rigid supervised AI for the UX workflow.

The AI assistance for both tasks has a summarising aspect, but its exact function and how it is used is part of what will be investigated. Since all the content is in Dutch, the AI assistance is naturally also in Dutch.

#### Obslogging task

For the obslogging tasks, the AI assistance is given in the form of descriptive labels, one per sheet row. These labels are based on the 'Observation/Quote' column, which is filled in by the participant. This way, the user input helps generate the AI output, allowing users to guide it to some extent. The prompt given to the AI model to produce these labels is as follows

*“Please generate 1 unique and summarising label of at most 3 words, based on the following observation or quote:*

*(Genereer alsjeblieft 1 uniek samenvattend label van maximaal 3 woorden, gebaseerd op de volgende observatie of citaat:)*”

The length of the labels was kept short (between one to three words) to 1) imitate the format of their manual labels (i.e. 'Label', column E), 2) to keep the extra information that the researcher has to process to a minimum, and 3) to not take up too much screen space. Additionally, the role the system is assigned is a UX research assistant, who helps label observations and quotes based on their content. The temperature used for the model is 0.7.

The AI output is displayed in its own column, labelled 'AI Label' (see Figure 4.1) as soon as the participant has finished typing an observation or quote in the Jth column (with



several seconds delay). The AI assistance for obslogging is thus fully automated and **real-time**. This is achieved using Google App script. With this method, there is no possibility of evaluating and adjusting the AI output before it is displayed to the participant. However, it was deemed infeasible to wizard-of-oz the assistance during the obslogging, since the exact observations are inputted real-time and cannot be predicted beforehand. Nevertheless, it will be insightful to see how the participants react to possibly flawed AI output.

Initially, the idea of AI labels was more similar to the ‘Task’, ‘Context’ and ‘Label’ labels, which were frequently described as taking too much effort and a struggle. However, the complexity and flexibility per project and person, as well as the evolving nature of the labels, made this difficult to envision how to do that with AI. Or at least, a pattern suggesting what prompts could be used or what kind of labels should be produced was not discovered within the experiment preparation phase. Hence, we eventually settled for a summarising type of label.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Klok-tijd (automatisch)	Automatisch verschijnt de klok-tijd naast de observatie. Die kun je vergelijken met de klok-tijd in beeld in de opname.				Context	Label	AI Label	Impact	Observatie of citaat	Vink	Inzicht
2		Let op! Als je de naam van dit werkblad of van de kolommen aanpast, dan werkt deze timestamp functie misschien niet meer.								Wil je alvast aangeven of iets goed of slecht gaat? Typ getallen 1-4 if "" in kolom H om impact aan te geven met een kleur.		
3												
4												
5			RI, Frits, 28 (Desktop)	pre-test	Add to cart					Kan de knop niet vinden	<input type="checkbox"/>	
6			RI, Frits, 28 (Desktop)	1. Filteren	product-detail pagina				1	"Oh wat leuk, dit is precies wat ik nodig had. Nou zeg, dat lost echt iets op."	<input checked="" type="checkbox"/>	
7	13:33:37		RI,	pre-test	achtergrond						<input type="checkbox"/>	
8											<input type="checkbox"/>	
9											<input type="checkbox"/>	

FIGURE 4.1: Obslog with AI assistance

## Debriefing task

For the debrief task, AI assistance is given in the form of a bulleted summary list of the interview (of one respondent) in the issue list, based on the observations and quotes taken from the original project’s Obslog document (see Figure 4.2). For the summary, a bulleted list format was chosen with one AI point per line for easier processing (due to segmentation). Additionally, the observation or quote used to produce the given point is given alongside it in between quotation marks, to give the participant some explanation regarding from what the given point originated. These will be called *AI references*, see column F in Fig. 4.2). The list is only based on one respondent because that is the same amount of content the researchers will be instructed to use for their debrief. The GPT 3.5-turbo model generated the AI output for the debrief task with a temperature of 0.7. The prompt given to the model was

*“Please generate 10 unique and summarising points based on the observations and quotes in between the ‘text’ XML tags. Support each point with a specific reference or quotation from the given text.*

*(Genereer alsjeblieft tien unieke samenvattende punten gebaseerd op de observaties en citaten die aangegeven worden met de XML tags voor ‘text’. Ondersteun elk punt met 1 of 2 specifieke referenties of citaten uit de geleverde tekst.)”*

More requirements given were *“Every point is a phrase of 10 or fewer words; each point is unique. (Elk punt is een zin van tien of minder woorden. Elk punt is uniek.)”* This

constraint was given to keep the summary short and concise, and limit the number of words to be processed by the participant.

Slight adjustments (e.g. adding more context or role) to the prompt were attempted for all the respondents within the limited time, but the summaries with the previously outlined prompt of two respondents were eventually used. The two summaries were selected by comparing the AI-generated output with the original project report, and checking how many topics or points aligned and whether the points made sense and were understandable. The AI-generated summary is given to the participants before the debrief (preparations) by manually inputting them in the blue row. It was decided to manually add the AI Debrief output, to be able to review the AI output and select a more satisfactory summary that somewhat overlaps (in topics) with the actual project report. This was deemed necessary to be able to obtain more valuable insights (where at least part of the AI output could be somewhat correct and appropriate). As well as being able to compare between participants due to consistent AI-generated assistance. It was decided to display all ten points within one cell, to not take up the whole screen and leave space for researchers to write their own notes. These points will be called AI points (see column E in Fig. 4.2)

Impact	Wie schrijft?	Geschreven?	Nagelezen?	Inzicht titel	Quotes en/of Next actions
	AI			<ol style="list-style-type: none"> <li>1. Er is te veel content, waardoor het een blur wordt.</li> <li>2. Lange teksten kunnen leiden tot afhaken.</li> <li>3. Duidelijkere kopjes zijn gewenst voor betere navigatie.</li> <li>4. Er is behoefte aan interactie en meer visuele aantrekkelijkheid.</li> <li>5. Een app heeft de voorkeur boven een website.</li> <li>6. Een account heeft weinig toegevoegde waarde.</li> <li>7. Nieuwsbrieven lijken te veel op elkaar qua onderwerpen.</li> <li>8. Een account biedt meer dan alleen filteren van informatie en inspiratie zijn beide gewenst.</li> <li>9. Filteren van informatie en inspiratie zijn beide gewenst.</li> <li>10. De leeslijst wordt zelden teruggekeken.</li> </ol>	<ol style="list-style-type: none"> <li>1. "Op moment dat je zwanger wordt dan krijg je zoveel content dat alles een beetje een blur is."</li> <li>2. "Tekst te lang, zou afhaken."</li> <li>3. "Ik wil duidelijkere kopjes, nu begint alles bij 14 weken, dat weet ik nu wel."</li> <li>4. "Ik wil altijd een beetje interactie, ik wou dat die kiwi groter was want nu voelt het meer als huiswerk."</li> <li>5. "Dit is visueel aantrekkelijk en minder zenderig, meer aandacht aan de kopjes besteed."</li> <li>6. "Ziet geen meerwaarde in een account."</li> <li>7. "Alle nieuwsbrieven lijken een beetje op elkaar ook qua onderwerpen."</li> <li>8. "Ik had wel willen weten dat een account meer is dan alleen filteren van informatie en inspiratie zijn beide gewenst."</li> <li>9. "Het is fijn als er dingen voor je gefilterd kunnen worden maar inspiratie is fijn beetje 50/50."</li> <li>10. "Ik kijk nooit meer terug naar de leeslijst."</li> </ol>

FIGURE 4.2: Issue list with AI assistance: AI points in column E cell; AI references in column F cell. \*Client names are removed for GDPR reasons

#### 4.2.2 Dependent variable for RQ1 & RQ2

Since we want to explore the impact of AI assistance on the UX testing workflow, we don't want to limit the findings by putting major constraints on what we want to observe or hear about experiences. Hence, we begin observing and asking with a broad scope, before narrowing down to notable observation themes and asking in more detail based on what is being said.

For observations (RQ1), since we are interested in comparing the with and without AI condition, attention is paid to differences between UX researchers' standard approach versus the approach with AI assistance (including usage of AI assistance). Possible aspects that could be observed include how the columns are used for obslogging, what actions participants perform to prepare for debriefing, how they structure their issue list, etc. Moreover, attention is paid to actions during the with AI assistance test condition concerning ethical risks.

For experiences (RQ2), as said we start with generally asking how the participants' experience was (with and without AI assistance). However, we will also specifically ask about the content and presentation of the AI assistance, to get more insights into the why behind participants' experiences (as well as asking the reason behind any of the participants' statements).

### 4.2.3 Dependent variable for RQ3: ECL & GCL

Since we are interested in the effect of AI textual data filtering assistance on the cognitive load of UX researchers, what we want to measure is the extraneous CL. Additionally, since the tasks include/contribute to the analysis, which is an important step to reach the project goal, we also want to measure germane CL.

Since we are interested in how UX researchers experience the addition of AI assistance, a subjective measure in the form of a questionnaire is utilised to measure cognitive load. The questionnaire is based on the Paas and Leppink CL measures, which both have been evaluated (Leppink et al., 2013; Paas, 1992) and are highly cited. There is still critique on both measures regarding the reliability and validity of self-evaluation. However, many of the objective measurements require a physical experimental set-up, whereas the set-up for this research was decided to be online for easier video recording, including screen recording. Plus, for our research, we focus on participants' experiences, which are often subjective. Hence, a combination of the two self-evaluation questionnaires was deemed the best and most feasible measurement option.

The Paas measure is included as an addition to the Leppink because it can possibly provide a more general indication of mental effort, and it only consists of one question. The Leppink questionnaire was adapted to the context of UX testing, but the formulation was kept as similar as possible to retain the scoring validity. Nevertheless, the sentence structure was slightly changed to try to make it less biased towards an extreme of the descriptor used. For example, one of the original Leppink ECL questions was: *The instructions and/or explanations during the activity were very unclear*, which had to be rated from (1) 'not at all the case' to (10) 'completely the case'. How this question is formulated is skewed towards the instructions being very unclear. Therefore the formulation was changed to *The instructions and/or explanations during the activity were*, which had to be rated from (1) 'not unclear at all' to (10) 'very unclear'. Furthermore, with the pilot study, it was discovered that an ECL question was too confusing due to the double negative, resulting in an unrepresentative answer. Hence the two ECL questions were adjusted, e.g. from *helemaal niet onduidelijk* (not unclear at all) to *helemaal niet duidelijk* (not clear at all) and *ontzettend onduidelijk* (very unclear) to *ontzettend duidelijk* (very clear). Several items were excluded, due to having no relevance in this study's context. Furthermore, for each rating question, a corresponding open question was added, where participants can provide additional clarification should they feel it's needed. Doing so, we hope to collect more information on the reasoning behind certain ratings.

See appendix B for the questionnaires. Slightly different questionnaire is used for the two tasks, to account for the different tools used (Obslog and issue list). Lastly, the questionnaires were given immediately after each task was completed to prevent participants from forgetting what occurred as much as possible.

### 4.2.4 Controlled variables

Several variables will be controlled, to limit possible influences on the results. The controlled variables are summed up in Table 4.1.

Controlled variable	Control
Order effect of a within-subject design	Design control: randomise the order of the two test conditions
Prior knowledge of the project	Design control: exclude UX researchers who produced the project data that is used
Expertise with UX testing	Design control: Varied years of experience between participants
ICL / Task (project) complexity	Design/statistical control: Use same project data for all respondents + measure on a scale how difficult or complex respondents found the task/project (topic) “Essential cognitive processing; refers to the complexity or difficulties inherent in the to-be-processed material”
Work style	Design control: Having multiple participants

TABLE 4.1: Controlled variables and how to control them

### 4.3 Test day simulation

It was decided to simulate a test day rather than conducting the experiments during an actual project to 1) have more control over what happens, 2) to prepare and check the debrief AI assistance and 3) to not burden the researchers during their work. Moreover, it was decided to only use two interviews/respondents, because more will make the experiment session even longer when the current set-up is already planned to take four hours. Since the UX researchers also have limited time available for such projects, the current planning already takes the maximum possible amount of time.

A shortened version of a UX testing test day was simulated using interview videos of a previous project. Since UX researchers usually conduct preparations themselves, which helps with familiarising themselves with the topic and project, participants are given the test script as preparation, at the end of the recruitment process.

For the tasks of obslogging and debriefing, the participants are instructed to perform them as they usually would for the without AI assistance condition. For the with AI assistance condition, they are informed of the AI assistance and its functionality but are told to use it however they want, and that their task objective is still the same. For the Debrief task, participants are given preparation time if needed, since it is common for UX researchers to prepare to some degree during a test day.

The experiment is conducted online (via Google Meet) to make it easier for more participants to participate, as well as accurately represent their actual work environment when conducting online UX tests.

### 4.4 Materials

Previous UX testing project data is used, including videos of two interviews that the participants will watch during the first task, and obslog and report documents as input and guidance to create the labels and summaries for the wizard-of-oz AI textual data filtering assistance. The available projects were limited due to GDPR legislation and suitable content. Moreover, due to the restrictions on data retention of three months, only

the data of one project was available for the desired time scope/planning. This should be taken into consideration when evaluating the results of the debrief discussion task, as discussing the same project twice in a row might affect the task difficulty.

## 4.5 Data analysis

### 4.5.1 Data collection

During the experiment, several data materials are collected: the obslog notes produced by the participant, the CL questionnaires (two versions), the screen recordings during the obslog and discussion tasks (and filling in the questionnaire), and video recordings of the interview. The screen recordings of the tasks are used for later observation and task analysis. The interview recordings are transcribed and analysed. The questionnaires are processed and used to perform statistical analysis.

### 4.5.2 Statistical analysis

The questionnaires' ratings are the quantitative results of the two test conditions that are collected during the experiment. Since we are interested in how the cognitive load of a task differs between them, we will perform a pairwise t-test to assess for significance. Beforehand, a test of normality (Shapiro-Wilk) will be performed to see whether a parametric or non-parametric t-test should be employed. JASP<sup>1</sup>, a statistical analysis program, will be used to perform the statistical tests.

### 4.5.3 Qualitative analysis

For RQ1, we compare observations regarding participants' approach to the task with and without AI assistance. For RQ2, we examine any comments/statements (from the transcripts) that can say something about their experience with AI assistance, including thoughts, feelings, opinions, etc. For both research questions, qualitative analysis will be performed through affinity diagramming in an online whiteboard. Any relevant observations and statements are collected per respondent and task, before trying to group them based on a common theme or link.

#### Chapter conclusion

Thus, to answer our research question we conduct experiments during which the participants conduct the tasks of obslogging and debriefing on a simulated test day, both with and without AI assistance. The AI assistance is created using the OpenAI API and incorporated into the thesis company's standard tools. For the Obslog task, participants will receive AI-generated labels based on the observations and quotes they typed up. For the Debrief task, the participants are given an AI-generated bulleted list summarising the interview. The obtained results will consist of the researcher's observations, the participants' statements (especially statements regarding their experiences and usage of the tasks' tools) and the filled-in cognitive load questionnaires. The results will be analysed using statistical tests and affinity diagramming.

In the next chapter, we outline the exact procedures of the experiments.

---

<sup>1</sup>JASP Team. (2024). JASP (Version 0.18.3)[Computer software]. <https://jasp-stats.org/>

## Chapter 5

# Experimental procedures

To give the reader a better understanding of the exact experimental procedures, this chapter outlines the exact process of the experimental sessions, as well as the proceedings of the pilot study conducted beforehand.

### 5.1 Pilot study

A pilot study was conducted in which one participant (a UX researcher who was excluded from the participant pool) performed the simulated test day during which the test condition with AI assistance was tested. During the pilot study, we checked whether the AI assistance operated smoothly; i.e. if AI output actually appeared in the Obslog, if it didn't take more than a minute, etc. Plus, the questionnaire and interview questions were evaluated. It was discovered that the double negation in two ECL questions impeded their comprehension, resulting in incorrect answers. In response, we modified the scale terms of the ECL items in an attempt to reduce possible confusion by removing the double negation. For the actual experiments, the test script will be given before the experiment day, allowing the participants to prepare, which better represents the workflow of an actual UX testing project.

### 5.2 Experiment

Before each experiment session, the participants received the test script of the project with the instructions to read as preparations for the experiment. During recruitment, they also received the informed consent form, detailing what the experiment would entail. With all the participants who signed the consent form, the following experimental procedures were performed (see Figure 5.1 for a visual outline). To start the experiment session, the researcher gave a short introduction to reiterate the content and goal of the experiment, and what the participants could roughly expect. They were informed that two variants of an obslog had been prepared, as well as two interview videos. Moreover, the participants were reminded that they were not evaluated for performing the tasks and that they could stop or ask questions at any moment. From that moment on, the experiment was video and audio-recorded.

In the preliminary phase of the experiment, the participants were asked background questions regarding their age, work experience (specifically concerning UX testing using the company's method), experience with AI usage and expectations of AI assistance during

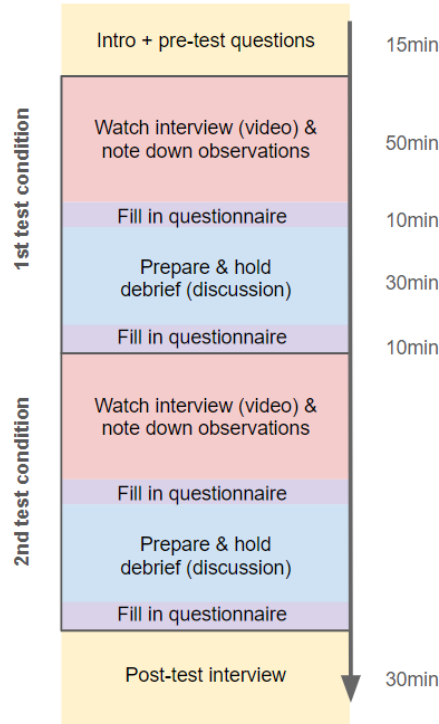


FIGURE 5.1: Experiment procedure & timeline

the tasks of obslogging and debriefing.

The experiment session consisted of two rounds, where each round has either test condition A (with AI assistance) or B (without AI assistance), which corresponds with the variants of the obslog documents. The number of times each test condition is applied in the first round is equally divided over the experiments; see Figure 5.2.

Participants are instructed to set up their work and desktop set-up how they usually would during an online UX test day, pretending the video is a live stream of the interview. Accordingly, they are also instructed to not pause the video.

During both rounds, the participants are asked to share the screen with the obslog document, which will then also be recorded. In each round, the participant first watches one of the interview videos whilst obslogging (logging observations and quotes, and labelling them). Next, the participant is instructed to hold a debrief (a discussion with the client concerning the highlights of the interview). Beforehand, they are informed to pretend that the interview is representative of a whole test day and that the researcher is the client. Moreover, the participants are given 5-15 minutes for debrief preparation in case the participant indicates that they usually do this. The exact amount of preparation time is based on the participants' judgement.

In the case of test condition A, the participants are briefed on what the AI assistance entails and briefly how it works. For example, for the obslog task, they are told that the 'Observations/quotes' in column J is used as input for the AI output (see Fig 4.1). For the debrief task, if they have not noticed it yet, they are pointed to the AI-generated summary with corresponding quotes in the 'issue list' sheet tab. For the test condition without AI assistance, the participants are reminded to perform the tasks how they normally would. For the condition with AI assistance, they are told to do whatever they want with the AI assistance and that the goal is still to execute the task to the best of their ability.

After each task, the participants are tasked to fill in the cognitive load questionnaire that corresponds to the task.

To control several variables, such as the order effect, the order of the test conditions and videos are randomised. Below the employed roster can be seen, where the last two columns refer to the experiment round, 'R' corresponds to the specific interview video used and 'a' or 'b' refers to the test condition.

<b>P#</b>	<b>1</b>	<b>2</b>
1	b + R2	a + R5
2	a + R2	b + R5
3	a + R5	b + R2
4	b + R5	a + R2
5	b + R2	a + R5
6	a + R5	b + R2
7	a + R2	b + R5
8	b + R5	a + R2
9	b + R2	a + R5

FIGURE 5.2: Participant timetable

Lastly, the participants are interviewed and asked about their experiences, impressions and thoughts regarding performing the tasks with and without AI assistance. The content and presentation of the AI assistance (output) are specifically touched upon. We also inquired about their reasoning behind their ratings of the questionnaire items and regarding any interesting observations made during the two experiment rounds. Furthermore, we also inquired about how/to what extent the presented AI assistance matched their expectations, and about any possible or desired improvements and/or changes to the AI assistance. After the interview, the participants are debriefed on how the AI assistance works, what will be done with the data and results, etc. Any questions they might have were also answered.

The test script that the researcher used for the experiment sessions, which includes the procedures and question guidance, can be found in appendix [A](#).

### **Chapter conclusion**

During the experiments, the participants conducted two rounds of a simulated test day with and without AI assistance. Each round consisted of 1) watching and obslogging an interview, 2) filling out the CL questionnaire for the Obslog task, 3) preparing and holding a debriefing, and 4) repeating step 2 for the Debrief task. Moreover, at the beginning and end of the experiments, participants were asked questions.

The next two chapters show the results obtained with these experimental procedures.



## Chapter 6

# Qualitative results

In this chapter, the qualitative results concerning AI assistance for obslogging and debriefing are given. For each task, observations made on the usage of the AI assistance and statements from the participants regarding their experience with the AI assistance are shown. First, however, we present an overview of the participants, their characteristics and whether they would use the AI Obslog and Debrief assistance in the future:

TABLE 6.1: Overview of the participants

P nr.	Position	UX testing experience	Age	Obslog AI assist.	Debrief AI assist.
P1	UX researcher (medior)	6 years	34	Would not use it	Would try using it during a real test day
P2	UX researcher (senior)	8 years	32	Would use it	Would use it
P3	UX designer (junior)	1.5 years	34	Would not use it	Would use it
P4	UX design intern	<1 year	22	Would not use it	Would maybe use it
P5	UX researcher (senior)	18 years	56	Would not use it	Would use it
P6	UX researcher (junior)	3 years	27	Would maybe use it if it was improved (e.g. more 'correct' and repeated labels)	Would use it
P7	UX research intern	<1 year	23	Would use it	Would not use it
P8	UX researcher (medior)	4 years	29	Would not use it	Would use it

## 6.1 Obslog assistance

This section first outlines participants' perceptions of the AI Obslog assistance. Then observations made and expressed experiences (or other statements) by the participants concerning the AI Obslog assistance will be outlined. Results are both regarding the usage of the AI labels for the task at hand (obslogging) and for a later moment, such as preparing for the debrief.

### 6.1.1 View of AI obslog assistance

Half of the participants did not use the AI obslog assistance nor comment on its potential usage and functionality. The other half described their view of the AI Obslog assistance as:

1. Assistance and inspiration for *theme formulation* (P2)
2. A small *summary* (of the 'Observation or quote' column) for quicker scanning (P3, P6-7)

### 6.1.2 Usage of AI assistance

This subsection outlines the observed and/or described usage of the AI assistance, during and after the task of obslogging. For the latter, the usage of the AI output for *theme formulation* and *a quick scan* will be presented.

No changes in how the UX researchers logged observations with and without AI assistance were observed. This is in line with how the participants stated that they did not use the AI labels for logging observations, or even looked at it during the task. In regards to the reasons behind this, P4 expressed how the AI assistance does not change the task of obslogging, thus not changing how intense and tiring the task is. Furthermore, one participant described the challenge of the AI output being displayed in real-time:

P6: "I won't read it while I'm writing. I would look at it afterwards (after completing the Obslog task)"

#### Theme formulation

For the usage for theme formulation, P2 described that when processing the interview content (e.g. reviewing the Obslog after finishing obslogging, during preparation for the debrief or when writing the report) some AI labels provided a nicely formulated and summarising term for the point or theme they were thinking of.

P2: "The AI assistance could work like a partner who thinks along with you; like a colleague who uses a specific word of which you think, 'Yes, let's use that'."

However, no concrete observations were made of this happening for P2, i.e. no AI label terms were found in the Issue list. Moreover, P5 expressed that the AI labels consisted of terms they would never use.

#### Quick scan

Several participants expressed how the AI labels could help to quickly scan the Obslog during later analysis by providing a small summary of the observation/quote cells. P3 only

expressed the possibility of using the AI labels but did not utilise them while preparing for the debrief. P6 and P7 stated they utilised the AI labels during the debrief preparations. P6 described how they first started preparing for the debrief using their normal approach, namely a ‘top-of-mind dump’. Next, they scanned the Obslog using the AI labels to find good, missing points to add to their initial list in the ‘Issue list’. P6 explains in the following quote how they would use the AI labels, to more quickly find the relevant ones to fully read.

P6: (About the AI Obslog assistance) “In one word, it shows the main message of the observation/quote, of her story. [...] Then you can first read the AI label, and if I think that I want to know more about it, then I’ll read this (the observation/quote).”

P7 mentioned that they typically review the Obslog, complete the ‘Impact’ column, extract important quotes, and form a mental overview of the most relevant points. They usually do this after finishing obslogging, often in preparation for the debrief. The participant described that they used the AI labels in combination with the self-written ‘Task’, ‘Context’, and ‘Label’ labels, to more quickly scan the Obslog.

Thus, a fourth of the participants (P6-7) used the AI Obslog assistance as a quick scan of the Obslog during Debrief preparations.

### 6.1.3 Experience with the AI assistance

This subsection presents participants’ statements related to their experience with the AI Obslog assistance. First, we summarise some general comments made, before showing more statements of the UX researchers grouped in themes of distractiveness and efficiency.

The majority of the participants viewed their experience with the AI Obslog assistance negatively. Many participants (n=6; P1, 3-6, 8) stated that the AI assistance provided no additional value for the task at hand, namely obslogging. Several (n=5; P1, 3-5, 8) also expressed that they found the assistance neither useful nor adding any additional value for later analysis. On the other hand, three participants (P2, P6-7) experienced the AI Obslog assistance more positively, describing it to be nice, useful and handy for later analysis.

#### Distracting

Several participants (P1, P6, P8) described the AI Obslog assistance as distracting because it took away their attention from the task of obslogging. This is because it changes what is on the participants’ screens, automatically attracting their eyes. Plus, they immediately mentally evaluate the AI output, distracting their brain. P8 comprehensively describes how the AI labels distract them:

P8: “I am easily distracted if something happens on my screen, so I have everything (like pop-ups) turned off. If I suddenly see some words move then I immediately think about it, I thought ‘oh yes, I agree or no that’s definitely not right’. But then my mind was on the AI labels and not on the interview/task at hand, and then I might have missed a relevant quote or observation. So at that moment, during typing (logging observations). I found it very annoying and disturbing.”

P7 remarked that it was subtle enough for them to ignore the labels and not use them when the AI labels were incorrect.

Thus more participants experienced the AI labels as distracting than those who found them subtle.

### **Efficiency**

The participants commented on various variables that are connected to efficiency. The variables mentioned are correctness, trust, checking the AI output, generating own output and high variance of labels. These variables are mostly mentioned in relation to time and/or ease of task procedures, thus concerning efficiency.

First of all, regarding *correctness*, two participants (P6-7) stated that the AI Obslog assistance (P6: sometimes) did a good job at summarising what they typed in the observation/quote cells in a few keywords. P6 then elaborated on how the correctness of the AI labels influenced their *trust* in the AI assistance and their future usage of it:

P6: “Sometimes he (AI Obslog assistance) summarises it (P6’s observations/quotes) really well [...]. That he writes it in two words is very handy. I think when I can trust that the AI assistance can do this well, it could be handy.”  
“If the labels are mostly pretty accurate, then I think I could really use them. Because then it will be very easy to scan (the Obslog). Then I don’t have to scan all my observations to derive my conclusions.”

So, sometimes the AI Obslog assistance produces good summarising labels. If the AI assistance is reliable in producing correct labels, it could save time and effort of not having to scan or fully read all their observations to make conclusions.

On the other hand, P3 and P8 commented how incorrect labels and not having written the AI labels themselves result in having to *check the AI output*, costing time and effort.

P3: “The AI labels should all be correct to be able to use them. Because else you will have to check every time, is the label even correct? And then you’ll be doing double (checking) work, and you’ll maybe be better off just checking out the observations.”

P8: “Because I did not type them (AI labels) myself, it makes (processing) it a little different. Then I have to look at it more thoroughly to see what it is about.

Lastly, P8 also criticised how the *high variance of AI labels* also contributed to additional checking of the AI output.

P8: “(Talking about the many different labels) Then I would have to read the whole column to understand and process the all. [...] Yes, looking at it (the AI Obslog assistance) again, the labels don’t have any additional value for me at this moment. This is because there are so many different labels, still resulting in a hundred separate items.”

Thus, the high variance of the AI labels leads to checking an increased amount of AI output, costing more time and effort.

## 6.2 Debrief assistance

In this section, we first provide participants' perception of the AI Debrief assistance. Then observations made and expressed experiences (or other statements) by the participants concerning the AI Debrief assistance will be outlined.

### 6.2.1 View of AI debrief assistance

The participants described the AI Debrief assistance as a...

1. Checklist; with reminders, suggestions (P1-3, P5-7)
2. Starting point (P2, P5, P8)
3. Common thread, guideline (P2, P4, P6)
4. Summary; overview of interview highlights (most important, relevant points) (P7-8)
5. Partner with own input, to help with 'afstemmen': similar to how 2 researchers help each other, to fill in each other; as reassurance (P3)

Many participants viewed the AI Debrief assistance as a checklist for missing, relevant points. As P1 says:

"I can imagine that it is interesting; not to just copy-paste, but to *check* if the AI mentions something interesting that I missed". (Would read it before the Debrief starts)

Moreover, P5 explains that sometimes missing relevant points is imminent.

P5: "Looking at it now (the AI Debrief assistance) it matches my points quite well... It contains a few weird ones, but I can imagine that you can nicely compare it with your own points. Like, have I seen everything? Is it complete? Because there will be a moment (a test day) where you'll forget a certain aspect or point."

The points provided in such a 'checklist' can be viewed both as suggestions and reminders, where the latter indicates that the point was not completely forgotten or missed, but merely not written down yet. P6 stated:

"I thought it was pretty cool that he (the AI Debrief assistance) could give *suggestions* in the issue list. Because sometimes I'm quite overwhelmed with all the obtained information, and then it is quite nice if he provides suggestions."

And P7 describes:

"I thought the first few points he (the AI Debrief assistance) gave were very useful. They were correct and accurate, and for me they were a nice *reminder*."

Other participants viewed the AI Debrief assistance slightly differently, namely as a starting point or guideline. When used as a starting point, the AI assistance is employed earlier in the task or process, rather than more towards the end when used as a checklist. P2 expressed that the AI Debrief assistance "*Gave direct, relevant input to structure the Debrief.*" P5 outlined usage as a starting point in more detail:

P5: “I would use this as a first *starting point*. I think I could use, or copy five or six points.”

P5’s words also indicate that further iterations of the issue list and its points will be performed.

Another way the participants describe the AI Debrief assistance is as an overview of the highlights (most relevant insights) or a summary of the most important interview content. This view is somewhat similar to a checklist and a starting point, in the sense that a list of points are given that can be used as input. However, a summary has a greater focus that all the main points of the interview are provided or that a broader view of the interview is given. P7 stated:

“Nice and clear list of insights that do a pretty good job at *summarising* what has been said”

P8 further described how the summary helped them:

P8: “All in all it’s a good *summary*. It helps you zoom out and write down overarching insights.”

Moreover, rather than using the AI output to start the issue list or fill in any missing gaps, P7 describes how they used it to quickly review what the respondent has said.

P7: “I thought it was nice to have a quick review, an overview of what the respondent had said and thought.”

Several participants (P2, P4, P6) viewed the AI output for the Debrief task as a guideline or a common thread. This perception is similar to a summary, where the emphasis lies on the main points of the interview content, but using it somewhat like a checklist, guiding the creation of the issue list. P2 explains how they use the AI Debrief assistance as a guideline:

“Actually, it is very nice that there is already some kind of *guideline* here. It would be especially handy if you would get this list for each respondent, each interview! Then you could scan on the report-making day and check if there’s a *common thread* or any outliers within the respondents.”

Moreover, it was observed that P4 used the AI output as a guideline since they did not add their own points, indicating that they thought the AI-generated issue list already contained all the main points necessary for the debriefing.

Lastly, one participant compared the AI Debrief assistance to a partner. The UX researchers at the thesis company usually work with a partner, who also writes down their findings and provides a second pair of eyes to check the issue list. P3 comments:

“It could be nice for when you’re checking the points written in the issue list. It’s kind of like there is a second UX researcher who noted down their own findings. Because sometimes you miss some points that could be relevant, and then I’ll think like oh yes, that’s a good one (referring to a point mentioned by their partner). It’s as if you’re not on your own, but that there is someone else who also watched the interview and can provide input for the Debrief.”

Thus, various statements from participants were obtained, either explicitly expressing a specific view and functionality of the AI assistance or describing it. Some participants used several descriptions to express themselves.

### 6.2.2 Usage of AI Debrief assistance

The exact approach and extent of usage varied between the participants, which was either observed or described by the participants themselves. For ease of presentation, we have organised the participants into three groups according to their AI usage. Group 1 (P1, P2, P5, P7) did not use it in a visibly observed manner and only expressed hypothetical usage. Group 2 (P3, P6, P8) employed the AI “Debrief” assistance as an addition to their standard approach to preparing for debriefing. Within group 2, there is also a distinction between those (P3, P6) who first focus on getting their own output on paper vs P8 who immediately focused on the AI output. Group 3 (P4) utilised the assistance to the greatest extent as a guideline during the debrief, which completely differed from their approach without AI assistance. Moreover, additional quotes are provided regarding the evaluation of the AI-generated points, the AI references (the observation or quote the AI point is sourced from) and the ethical risks concerning using the AI Debrief assistance.

#### Group 1: hypothetical or non-usage

For the following participants (P1, P2, P5, P7) actual usage of the AI Debrief assistance was not directly observed. Moreover, relevant statements regarding the usage of the AI output contain hypothetical word formulation, using ‘imagine’, ‘could’, ‘would’, etc. For example,

P1: “I can *imagine* that it is interesting; not to just copy-paste, but to check if the AI mentions something interesting that I missed, like oeh that’s interesting, I’ll include that (in the issue list and debrief)’ [...] I *would* be happy to read it (the AI list of points) before the debrief starts.”

and

P5: “I *could imagine*, that if I look at this (the AI Debrief output), that you can use it to check for any points you missed [...] So, I *would* use it as a first starting point.”

P2’s statements regarding the use of the AI Debrief assistance often did not contain hypothetical word formulation, however, their statements could be seen as contradictory. Take the following statements as an example:

P2: “It (the AI Debrief assistance) mostly confirmed things for me. I already had a clear view of the most important insights, so my understanding (of the insights and such) has not necessarily improved.”

and

P2: “I find it very effective. Usually, some points jump out for me, which I then write down (e.g. in the issue list). However, this respondent spoke very fast, so I didn’t manage to do so myself. It (the AI Debrief assistance) also extracted some nuances that I did not remember myself. And it had quotes to support my point.”

How these statements are contradictory will be discussed in Chapter 8.

## Group 2: addition to standard approach

Group 2 combined both AI output and output generated with the participants' standard approach, to prepare for the debrief. Within group 2, the approach can be further differentiated by the order of steps. P3 and P6 started with generating output in their usual manner before comparing that with the AI output. To illustrate,

P3: "I generated my own insights and now I'll compare it with what the AI produced."

and

P6: "I didn't literally copy-paste it. I just scanned the AI points whilst thinking 'what is this about again?' And then I supplemented what I had already written down (in the issue list) using the AI points, filling in any gaps. [...] It provides a couple of important insights and insights that are less important. Then I select which ones I find important enough to write down in the issue list."

On the other hand, P8 began with reviewing the AI output and using some of it to write down points, before scanning the Obslog like they usually do.

It was observed that P3, after a brief look at the AI debrief output, started preparations with their usual approach without AI assistance: scrolling through the Obslog, ticking the 'Video' checkboxes, adding comments to the 'Insight' column and copying them to the "Issue list" with further additions. After having finished scanning the Obslog, the participant reviewed the AI output, adding any points deemed relevant to the existing list of points. The points were completely or only partially copied, or formulated differently; see the following excerpts taken from the issue list with AI assistance filled by P3 (+ an accompanying quote).

AI Debrief output: 9. Filtering of information and inspiration are both desired.

P3 issue list: Filtering of information and inspiration are both desired.

AI Debrief output: 4. There is a need for interaction and more visual appeal.

P3 issue list: There is a need for interaction.

In addition to this excerpt, P3 also commented the following:

P3: "Need more interaction. I already had the visual appeal, but I thought that point was very good. "

AI Debrief output: 10. The reading list is seldomly looked at.

P3 issue list: The reading list won't be used.

P6 stated that they started the debrief preparations with a top-of-mind dump in the 'Issue list' and then scanned the Obslog. This was similar to the observed debrief preparation behaviour of P6 during the without AI test condition. Next, they reviewed the AI debrief list and expanded the issue list. Certain parts of the AI points were formulated differently, before including it as additions to their own already written down points. See the following quotes and issue list excerpt:

AI Debrief output: 6. An account has little additional value.

P6 issue list: (Under 'make account' header) Making an account goes well, but has little additional value.

Accompanying quote for the above excerpt:



P6: “For example, ‘an account has little additional value’ is a really good one; that is indeed important, so I added it (to their issue list)”.

During the without AI test condition, P8 performed debrief preparations by 1) starting with a top-of-mind dump, 2) simultaneously clustering and organising the written down points with headers (test script topics), 3) scanning the Obslog for missing items, particularly using the ‘Impact’ column for what to focus on, 4) editing and reorganising points written down. During the with AI test condition, P8 employed similar steps, but with the additional presence of the AI assistance. P8 started with checking out the AI output, whilst writing down points for the issue list and already organising them according to test script topics. Lastly, they scrolled through the Obslog to check for missing points. The following quote describes in more detail how the participant went about the debrief preparations and how they utilised the AI assistance:

P8: “In the end, I created my issue list with the help of the AI Debrief assistance. I examined the AI points, thinking about what each point is about, what is missing, etc. Next, using the AI points I filled in my issue list, but I did go through the Obslog again to check for any other, missing information to make the issue list more complete. Sometimes you have time for that, sometimes you don’t (to check the Obslog). In this case, I did have the time. [...] I did (often) miss what test script sections corresponded with the AI points, so I eventually went ahead and wrote those topics down myself.”

The following issue list excerpts show some overlap between the AI output and the participant’s own list, which were written before scanning the Obslog and whilst going through the AI list:

AI Debrief output: 1. The respondent uses multiple apps and mailing lists for pregnancy information. 2. The respondent looks for trustworthy sources and reads multiple articles for information.

P8 issue list: Platform | Apps/ mailing lists/newsletters; use multiple sources (added later, after scanning Obslog: and looks for familiar sources) but does not want it to become too much (information)

AI Debrief output: 3. The respondent appreciates visual content, like videos and visualisations.

P8 issue list: growth calendar | appreciates visual content, videos and visualisations

AI Debrief output: 4. The respondent has a need for clear and structured navigation, and summaries of the articles

P8 issue list: growth calendar | misses navigation

AI Debrief output: 5. The respondent finds transparency very important when making an account and sharing personal data.

P8 issue list: want more information on...

changed to: making an account should be more transparent.

The above issue list excerpts demonstrate some partial copying of the AI output and reformulation of certain parts/topics. P8 expressed some different attitudes about copy pasting the AI output at initial vs further glance.

P8: (First comment) “I thought, ‘Oh, I’ll go and copy-paste, and select points”  
(At a later moment) “There were a couple points that gave more emphasis on

certain parts that I did not find as important [...] So the AI list contained some stuff what made it that I couldn't copy the points completely as they were.

### **Group 3: guideline**

In the first round without AI assistance, P4 scrolled through the Obslog, writing down a list of points to be used for the debrief. However the second round with AI assistance, they disregarded this approach and utilised the AI output as their main guideline. P4 crossed out one point and highlighted specific parts of several points. P4 stated that they had never led a Debrief and were uncertain how it should exactly be done. Plus, they had never encountered the issue list tab.

Thus, various approaches and extents to the usage of the AI Debrief assistance were observed by the researcher and expressed by the participants. The usage extent ranges from hypothetical or non-usage to heavy usage as a major guideline. Next, we provide statements regarding participants' evaluation of the AI output.

### **Evaluation of AI points**

All participants were observed to evaluate the AI output, regardless of their usage extent, but to different extents. Here we collect general comments and evaluative statements per point.

Participants made general comments regarding their overall evaluation of the AI output, what they were missing and what that meant for how they completed the Debrief task. See the following quotes for illustration:

P4: (At first glance) "Looks to be very similar to what I wrote down the 1st round/ interview"

P2: "I did still have to mentally consider, think over the motivation -the why-behind the AI points. Now it (the AI points) says what happens, but not why it happens [...] I look at each point, and think what can I keep, change and discard?"

P6: "I still wanted to briefly check it myself, so then I scanned and reviewed the AI points."

In addition to participants' general comments, per point, evaluative statements were also given. These statements demonstrated that participants evaluated how 'correct' and relevant the AI points were, as well as whether the participants already had the point in their issue list. The correctness of an AI point is judged based on several aspects.

Firstly, participants assess how much of the content is true, in the sense that it matches what the respondent said or did (or at least how they remember the respondent's words and actions). For example,

P5: (Referring to the AI point: There is too much content, resulting in an overload.) "I don't think the respondent actually said there was too much content. They stated that there was too much information on one screen. I think that is different from there being too much content."

Secondly, participants also evaluate whether the AI point's conclusion is correct. To illustrate, we provide an example below where the participant finds the conclusion correct and another where they deem the AI point's conclusion incorrect.

P5: “(Referring to the AI point: Account had little additional value) I think this is definitely a correct and valid conclusion. So that point is correct. The respondent gave various reasons for it.”

P5: “(Referring to the AI point: an app is preferred over a website) I did not hear the respondent say this. She said something about it but in a totally different manner. Hence, I find this a completely wrong insight.”

Thirdly, the participants assessed whether the interpretation of the AI points was correct. For instance,

P7: “(Referring to the AI point: 8. The respondent looks for the source and references to assess the trustworthiness of the magazine.) I don’t think this matches with what the respondent said. I don’t know for sure whether the AI interpreted this correctly. I think this refers to the fact that it belongs to a bigger organisation. And whether that influences the magazine’s trustworthiness or not, ehm...”

Lastly, the participants evaluate whether the nuance of the AI point is right, i.e. if the emphasis is put on the right aspect. Take the following comments as an example:

P8: “There were several (AI points) of which I thought that some parts had gotten more emphasis than I thought was necessary. For example, ‘the respondent looks for trustworthy sources and reads multiple articles to check for correctness’, is true. However, that the respondent is constantly looking for trustworthiness is not true. It is recognisable; not trustworthy. So, there are a couple of things which make it so that I cannot outright copy the AI points word for word.

and

P7: “I’m not sure whether this was the crux of the interview, of what the respondent said. I think it was more of a, ‘Oh, this is not necessary so I won’t tick the checkbox.’”

Participants’ evaluation of the correctness of an AI point is also dependent on the text formulation. Participants might find the exact terms used not suitable or would use a different wording. To illustrate this:

P5: “But (AI) point three is spot on. Yes, clearer headers are needed for better navigation. That is... I would maybe not use the term navigation, because that suggests going to different pages, but this was on the same page. I would adjust that. For the rest, this topic resonates well with what the respondent said.”

Then, regarding the evaluation of relevancy, see the following statements:

P3: “(Referring to the AI point: the app is preferred over a website) I found this a pretty good one. Although, it was mentioned more in passing.”

and

P6: “I felt like it did not always give the most important insights [...] It provides a couple of important insights and insights that are less important. Then I select which ones I find important enough to write down in the issue list.”

Furthermore, a statement of P7 demonstrates how participants also evaluated the AI output by comparing it to their own points:

P7: “(Referring to the AI point: The respondent appreciates visual content, like videos and visualisations.) I already listed this point. I think this point refers to the respondent’s need for quick and clear information.”

Thus, participants’ statements exhibit evaluation of AI output, including agreement and disagreement. As well as assessment concerning partial sections of points.

### **AI references**

Participants also commented on the AI references, which were the source observations/quotes of the AI points. The majority of the participants perceived these references as quotes from the respondent because they were all displayed between quotation marks. The following statements show participants’ appreciation for the idea of quotes supporting the AI points and how they would use them.

P2: “The addition of quotes is very nice, because the client often looks for evidence of what we claim as insights, and it helps illustrate the points. Then clients cannot say that it is merely the interpretation of the UX researcher [...] It would be ideal if you have multiple quotes for a point/ insight. Then you don’t have to go through the Obslog again.”

P6 had a similar opinion, although they thought the AI references would be used more intensely on the report-making day. P5 further emphasised how the provision of quotes was very practical as it would save a lot of time and effort (since the UX researchers often put respondents’ quotes in their reports for the client).

Only P8 noticed that not all references were quotes, they expressed:

“I quite quickly got the idea that they (the AI references) were not only quotes that I wrote down (created by the participant typing it in quotation marks in the Obslog) but that it also showed what the observations I wrote down as quotes.”

P8: “En daar had ik al vrij snel het idee dat het niet alleen maar quotes waren, als in niet alleen maar cellen die ik met aanhalingstekens heb, maar dat het ook quotes zijn van delen die ik gewoon als observatie heb opgeschreven.”

Furthermore, P6 did realise at a later moment, after examining the AI references for longer, that they are not (all) written by themselves, and was curious as to where they came from then.

Several participants were very positive about the AI references, but incorrectly perceived them all as quotes from respondents. Only two respondents eventually spotted that some of the references were observations, but mistakenly displayed between quotation marks.

### **Ethical Risks**

Participants also expressed some concerns or considerations regarding possible ethical risks. For example, P1 explicitly stated:

“I spot some risks if we just give (the client) this top 10 (the AI Debrief assistance).”

Moreover, P6 and P7 also expressed concerns about the AI output influencing the results, possibly skewing them or making them biased. P7 further elaborated on in what situations the risk would be higher. See the following statements:

P6: “I looked a little bit at this (the AI points), but I didn’t want to get too influenced by it. Hence, I first wrote down points in the issue list based on what I remembered myself.”

and

P7: “I found the first few AI points very useful; they were correct and accurate and served as a nice reminder. However, the subsequent points contained aspects that the AI interpreted incorrectly. If I so happen to be a little lax as a researcher, taking the AI point as the truth, then I’ll have incorrect results. And that can be dangerous. So yes, then I would rather not use it so you don’t have that danger.”

Thus, three participants explicitly expressed some concerns regarding the ethical risks of having biased or incorrect results due to using the AI Debrief assistance.

### **6.2.3 Experience with the AI assistance**

Many comments were made regarding the AI debrief assistance, but they are generally directly related to the usage or functionality of the AI assistance, so there is some overlap. Regardless, in this subsection, first, some general (positive) comments related to experience are given. Next other comments are grouped by themes of effectiveness & efficiency, correctness, and presentation of the assistance.

#### **General positive responses**

Generally, participants had a positive impression of the AI Debrief assistance, describing it as nice to have, finding it impressive, cool, surprisingly good and interesting. For illustration:

P2: “Very impressive! I think this is great, the AI assistance. I find it very cool to see.”

and

P8: “When I read the first few AI points, my first reaction was ‘wow, where? How? Where does this come from? Very cool.”

#### **Effectiveness & Efficiency**

Several participants (P2 and P5) directly expressed how they felt the AI assistance affected the effectiveness and efficiency of the task at hand.

P2: “The AI input really helped with getting the most important points from the interview, which saves time and effort.”

P5 elaborated on how they thought the AI assistance helped with efficiency, but not effectiveness. See the following quote:

P5: “I don’t think the AI made it more effective, but perhaps it did make it more efficient. For me, effectiveness refers to achieving what you want, and for

efficiency it is about how much effort it takes you. This AI assistance reduces the amount of effort needed.”

Although P5 stated that the AI assistance did not help with effectiveness, they did say it helped with making their issue list more complete by adding relevant points they missed.

P5: “Opnieuw, ik vond niet dat de AI hem effectiever maakte. Maar misschien wel efficiënter. Ik zit altijd met het verschil. Voor mij zijn dat twee hele verschillende termen. Het ene gaat over krijg je voor elkaar wat je wil. Nou, in allebei de gevallen. Maar bij efficiënt gaat het over hoeveel moeite doe je ervoor. En deze haalt wat moeite weg.”

“Ik ben nu iets vollediger, want hij pakt nog een paar dingen uit... die ik zelf misschien niet meteen erbij had gehaald... maar wel relevant vind achteraf.”

On the contrary, P1 commented that they thought the AI assistance would not have a big impact on their work, referring to the amount of time and effort it takes them to perform the Debrief and preparations for it.

### **Correctness**

Various participants made comments regarding their perceived correctness of the AI output, which overlaps with the earlier Section 6.2.2 on the evaluation of AI output. For example, P5 stated that some points were spot on. All the participants commented on at least one AI point, which they said was not (completely) correct and one point that they deemed correct.

### **Presentation**

Some participants also had concrete comments on the presentation of the AI assistance. For example,

P1: “Nice that point & ref are separated; don’t need headers (kopjes). I find it concise and powerful (short and sweet) like this.”

## **6.3 Experimental set-up**

During the experiments, several participants commented on the experimental set-up in terms of the simulated test day vs normal test day and the questionnaires. Commented differences between the simulated and normal test days are a lack of actual clients and a lower number of interviews and respondents during a test day (especially before conducting a debrief).

Several participants (P1, P5) commented on the lack of actual clients, resulting in the participants experiencing less pressure. Since clients’ input helps guide the debrief, the task approach during the experiment somewhat differed from an actual project. P1 further describes how real clients impact the debriefing:

P1: “(About the client) I just want to know whether they have different or wrong conclusions based on what they saw, which I would have to rectify (during the Debrief). If that is the case, I’ll have to put in more effort to go against it. Plus, it also shows what the client is more interested in, so what I should pay more attention to (when writing the report).”

P1 also explicitly stated how the simulated test day differed from a normal test day.

P1: “I thought it (the debriefing of the simulated test day) was very easy as I did not have an actual client I had to convince.”

Furthermore, several participants (P1-3, P5) commented on how insights, themes, and links between insights usually only emerge after multiple interviews (at least more than two). P1 also described how that affected their work procedures demonstrated during the experiment.

P1: “(Regarding the use of the issue list) Usually I do use it after I’ve done multiple interviews. When certain topics are mentioned more often; then I write them down in the issue list. And I would also write down what we just discussed (during the Debrief). I would fill in the issue list during quiet moments, for example after the second or third interview.

Lastly, we have some comments for the cognitive load questionnaires. For instance, several participants (P1-4, P8) also asked for clarification for the GCL questions, such as what exactly is meant by insights. Plus the improved understanding is seen as inherent to performing the tasks. For illustration see the quote below.

P3: “Obslogging improved my understanding of the obtained insights, a lot or not really. But the obtained insights refer to your own insights that emerge during obslogging. Isn’t that the principle of obslogging?”

Plus, P1 explained their CL ratings out loud:

P1: (for Q4, regarding the Obslog task with AI assistance) “I’ll give it a five, there was not a big difference with the previous one (the other test condition, without AI assistance).”

However, their actual ratings contradicted their statement. Examining the results of the cognitive load questionnaire, P1 gave an eight and a five for Q4: complexity of interview structure.

## 6.4 Obslog vs Debrief AI assistance

Since each participant conducted both the Obslog and Debrief task with AI assistance, several participants also compared the two, in general terms, but also regarding the amount of context provided for the AI output. Commentary regarding similarities between the AI assistance of the two tasks is also provided.

Generally, participants (P1-5, P8) were more positive regarding the AI Debrief assistance. As P5 describes:

“I think this (the AI Debrief assistance) is of a whole different (higher) level than what was given in that Obslog column (referring to the AI Obslog assistance).”

Concerning the context needed to understand the AI output, P7 details that they prefer the AI Obslog assistance over the AI Debrief assistance, because of the context in which it is displayed, as well as the consequences of this on the ease of checking the AI output.

P7: “I quite liked the AI labels. Like that I can quickly see the ‘Context’, ‘Label’ and AI labels. Those three together give me a comprehensive view of what has been said [...] However for the issue list, the AI points are removed

from their context. It just gives you ten points from the whole interview, making it difficult to check whether the point is true or not. Then I have to rely on my memory for verification. That's where it can go wrong."

P6 did not explicitly compare the amount of context provided for the AI assistance for either task, but did criticise the lack of context for the AI Debrief assistance.

P6: "For example here he (the AI Debrief assistance) says 'improved communication'. But without knowing where exactly in the interview this has been mentioned, then it is difficult to find it back to understand what it is actually about/ what it specifically refers to."

Besides the differences between the two tasks and the AI assistance provided for them, some similarities were also observed and commented on. It was observed that all participants were curious about the AI assistance provided, and many participants commented positively on the integration of the AI assistance in their standard tools.

All the participants were curious towards the AI assistance, often eager to try out the AI Obslog assistance by typing something and checking the issue list to see if the AI already outputted something.

Many participants (P1-2, P4-5, P7) commented positively on the integrated manner in which the AI assistance was provided. P1 mentioned it to decrease the barrier to using the AI assistance. P2 mentions familiarity and no need to spend time and effort to learn how to use a new tool or platform as benefits.

P2: "Instead of having a new tool, which would mean that I would have to do away with all the habits that I built up in the last five years, which make me so fast (in performing the test day tasks), because I want to work with AI. But now you have found a way to incorporate it into my workflow."

"When I got your invite I thought 'Oh, there we go again with the umpteenth ChatGPT tool, which I'll just put away in some list (like a bookmark for later use). The newest AI tool that is supposed to enrich my life, but actually gives extra noise."

### **Chapter conclusion**

In conclusion, a lot of qualitative results have been collected in the form of the researcher's observations and the participants' statements. The qualitative results have been organised according to the usage of and experience with the AI assistance for the two tasks of obslogging and debriefing. Next, they are grouped according to discovered themes such as the various views or exact usage of AI assistance, efficiency and effectiveness and more. Furthermore, we presented results concerning the experimental set-up and the differences and similarities between the AI assistance for the Obslog and Debrief tasks. These findings will be discussed in Chapter 8.

In addition to the results given in this chapter in the form of observations and statements, the next chapter provides more results concerning participants' experienced cognitive load in the form of quantitative data and additional clarifying statements of the participants.



# Chapter 7

## Cognitive load results

The cognitive load of participants during the Obslog and Debrief task, with versus without AI assistance was measured using questionnaires. This included questions concerning the general mental effort on the Paas scale (Q1), the extraneous cognitive load of participants (Q5-6) and the germane cognitive load (Q7-9). A more complete overview of what the questions entailed is given below. Each question has values for test conditions A: with AI assistance and B: without AI assistance. These values are indicated by a term composed of the question number + the test condition, e.g. Q1A refers to the rating value for Q1 with test condition A (with AI assistance). The gathered results are displayed in tabular (and graphical) format per task, along with a summary of participants' additional clarification provided grouped by ME, ECL and GCL questions.

The questionnaire questions are labelled numerically in the result tables, so here is an overview of what each question entails:

Q1 Mental effort (Paas scale)

ME: A negative difference between A and B represents a reduction in effort when performing the task with AI assistance (compared to without).

Q5 ECL: clarity of the roles (functionality) of the task's tool

Q6 ECL: effectiveness of the task's tool for the task at hand

ECL: A negative difference between A and B represents a decrease in clarity or effectiveness of the task's tool when conducting the task with AI assistance (compared to without AI assistance).

Q7 GCL: improved understanding of the project topic

Q8 GCL: improved understanding of the obtained insights

Q9 GCL: improved understanding of links between insights and corresponding themes

GCL: A negative difference between A and B represents a decrease in improved understanding after conducting the task with AI assistance (compared to without AI assistance).

## 7.1 Obslog task

For the Obslog task, all participants (n=8) filled in the cognitive load questionnaire. For the ratings, we computed the descriptive statistics including a test of normality (see Table 7.1), and a parametric and non-parametric pairwise t-test (see Tables 7.2, 7.3).

TABLE 7.1: Obslog: Descriptive statistics, n=8

	Mean	STDEV	Shapiro-Wilk	Shapiro-Wilk p
Q1A	5.500	1.690	0.814	0.041
Q1B	5.875	1.808	0.861	0.123
Q5A	7.125	2.232	0.887	0.220
Q5B	7.750	2.252	0.889	0.230
Q6A	7.875	1.553	0.952	0.731
Q6B	7.625	1.188	0.892	0.245
Q7A	7.500	1.512	0.918	0.416
Q7B	7.250	2.435	0.857	0.113
Q8A	7.625	1.302	0.877	0.178
Q8B	7.625	1.923	0.939	0.603
Q9A	6.500	0.756	0.724	0.004
Q9B	6.625	1.506	0.871	0.156

For all questions, except Q1 and Q9 for the test condition A (with AI assistance), the assumption check of normality (Shapiro-Wilk; shown in Table 7.1) is not significant ( $p > 0.05$ ) suggesting that the pairwise differences are normally distributed, therefore the assumption is not violated. Since the assumptions of normality for two rating distributions are violated, both the student's and Wilcoxon signed-rank t-tests have been performed (see Tables 7.2, 7.3).

TABLE 7.2: Obslog: Paired samples student's t-test

Measure 1	Measure 2	p	Mean difference	SE Difference
Q1A	Q1B	0.528	-0.375	0.565
Q5A	Q5B	0.388	-0.625	0.680
Q6A	Q6B	0.563	0.250	0.412
Q7A	Q7B	0.685	0.250	0.590
Q8A	Q8B	1.000	0.000	0.655
Q9A	Q9B	0.836	-0.125	0.581

TABLE 7.3: Obslog: Paired samples Wilcoxon signed-rank t-test

Measure 1	Measure 2	W	p	Hodges-Lehmann Estimate	Rank-Biserial Correlation
Q1A	Q1B	5.000	0.586	-0.500	-0.333
Q5A	Q5B	4.000	0.408	-1.000	-0.467
Q6A	Q6B	17.500	0.588	$4.728 \times 10^{-5}$	0.250
Q7A	Q7B	12.500	0.750	0.500	0.190
Q8A	Q8B	7.500	1.000	0.000	0.000
Q9A	Q9B	6.500	0.892	-0.500	-0.133

For both paired samples t-test and all the questions  $p > 0.05$ , thus the differences between the ratings (for mental effort, ECL and GCL) of conducting the Obslog task with and without AI assistance are not significant.

### 7.1.1 Mental effort (Paas scale)

The participants provided several factors that contributed to their mental effort ratings for the Obslog task, namely the nature of the task (P3-4; “the task of obslogging by itself requires a lot of focus and effort”), familiarity with the task and project topic (P2-3, P6; standard, 1st vs 2nd round) and talking speed of the interview respondent (P2-3, P5, P8) were most often mentioned. Only P4 and P5 mentioned the AI labels positively and negatively respectively, but both expressed that they only had a small impact.

### 7.1.2 ECL

In the additional explanations, P3 and P4 mentioned that the ‘Task’, ‘Context’ and ‘Label’ columns impacted how clear the roles (functionality) of the Obslog columns were. This is because of confusion regarding how to use those columns, because of the subjective and ambiguous division between them. P4 and P8 explicitly stated that the AI label’s role was clear, but also commented that it was distracting. Moreover, P5 and P6 expressed that they thought there were too many columns. P6 further explained that as a result, the AI labels were more difficult to get used to as they gave extra ‘noise’. Nevertheless, P6 stated they would then remove the ‘Label’ column rather than the AI labels. Regarding the effectiveness, the participants barely mentioned the AI labels in the additional clarification for their ratings. P4 and P8 do state that for them the AI assistance did not have additional value for obslogging, and did not make it easier than normal.

### 7.1.3 GCL

For all three rating questions, the variable most often mentioned by participants in the additional explanation is the nature of the task; understanding of everything improves by doing the task, hence improving whilst obslogging multiple interviews. Plus, P6 and P8 explain that the improved understanding of links between insights and corresponding themes mostly occurs during later analysis on the report-making day. For their ratings for the improved understanding of the obtained insights and the links between them, P2 and P6 mention the AI labels. P2 describes that the AI Obslog assistance helps with thinking of a suitable term for emerging topics and themes. P6 clarifies that the AI labels only sometimes help. This is because when the AI labels don’t summarise the input column well, P6 still has to expend extra time and mental effort scanning their observations and quotes.

## 7.2 Debrief task

For the Debrief task, all participants except for P7 (n=7) filled in the cognitive load questionnaire. P7 did not complete the questionnaire due to time constraints. For the ratings, we computed the descriptive statistics including the Shapiro-Wilk test of normality (see Table 7.4), and a parametric and non-parametric pairwise t-test (see Tables 7.5, 7.6).

For all questions, except Q8A, the assumption check of normality (Shapiro-Wilk; shown in Table 7.4) is not significant ( $p > 0.05$ ) suggesting that the pairwise differences are normally distributed, therefore the assumption is not violated. Since the assumption of normality

TABLE 7.4: Debrief: Descriptive statistics, n=7

	Mean	STDEV	Shapiro-Wilk	Shapiro-Wilk p
Q1A	4.286	1.799	0.893	0.292
Q1B	3.857	2.410	0.873	0.196
Q5A	8.857	1.069	0.894	0.294
Q5B	6.857	3.848	0.805	0.045
Q6A	8.000	1.633	0.933	0.573
Q6B	5.714	2.628	0.868	0.179
Q7A	6.286	1.604	0.880	0.224
Q7B	7.587	0.690	0.840	0.099
Q8A	7.286	1.704	0.795	0.036
Q8B	7.286	0.756	0.833	0.086
Q9A	6.143	1.215	0.859	0.147
Q9B	7.857	0.900	0.818	0.062

for one rating distribution is violated, both the student’s and Wilcoxon signed-rank t-tests have been performed (see Tables 7.5, 7.6).

TABLE 7.5: Debrief: Paired samples student’s t-test

Measure 1	Measure 2	p	Mean difference	SE Difference
Q1A	Q1B	0.667	0.429	0.948
Q5A	Q5B	0.162	2.000	1.254
Q6A	Q6B	0.098	2.286	1.169
Q7A	Q7B	0.033	-1.571	0.571
Q8A	Q8B	1.000	0.000	0.617
Q9A	Q9B	0.037	-1.714	0.644

TABLE 7.6: Debrief: Paired samples Wilcoxon signed-rank t-test

Measure 1	Measure 2	W	p	Hodges-Lehmann Estimate	Rank-Biserial Correlation
Q1A	Q1B	11.500	0.915	$6.090 \times 10^{-5}$	0.095
Q5A	Q5B	9.000	0.201	3.500	0.800
Q6A	Q6B	14.000	0.106	3.000	0.867
Q7A	Q7B	0.000	0.058	-2.000	-1.000
Q8A	Q8B	8.000	1.000	$1.063 \times 10^{-5}$	0.067
Q9A	Q9B	0.000	0.057	-2.500	-1.000

For both paired samples t-tests and all the questions except for the student’s t-test of the GCL Q7 and Q9  $p > 0.05$ , suggesting that the differences between the ratings (for mental effort, the ECL items and GCL’s Q8) of conducting the Debrief task with and without AI assistance are not significant.

The p-values of the paired samples student’s t-test for Q7 and Q9 are  $< 0.05$ . However, the p-values of the paired samples Wilcoxon t-test are all  $> 0.05$ . We look at the additional clarification given for the GCL questions to evaluate these contrasting values in Section 7.2.3.

### 7.2.1 Mental effort (Paas scale)

All the participants who provided additional clarification mentioned AI assistance as a positive element for decreasing mental effort for the debriefing task. On the other hand, P2-4 also expressed how the AI assistance negatively impacted the mental effort ratings, because of the effort required for checking and editing the AI output, and “staying critical”. P2-3 and P8 also mention the nature of the task as a contribution to the mental effort expenditure for the Debrief task.

### 7.2.2 ECL

P2-3, P6 and P8 positively mention the AI output as being effective for executing the debriefing task. P4 mentions inexperience as the main factor contributing to their rating of the effectiveness of the issue list.

### 7.2.3 GCL

Similar to the explanations for the Obslog task, the nature of debriefing is mentioned several times as a contributor for GCL (Q7: P2-4, P8; Q8: P3; Q9: P2, P4).

#### Chapter conclusion

In conclusion, with the Shapiro-Wilk test, we found that the quantitative results for almost all the CL questions had a normal distribution. Nonetheless, since the assumption of normality was violated for some questions, we performed both a parametric and a non-parametric t-test. The results of the t-tests overall suggest that the differences in ratings for mental effort, the ECL and GCL items between the two test conditions were insignificant. These quantitative findings will be discussed in the next chapter with the help of participants’ additional clarifications for their ratings, presented in this chapter, and the qualitative results from Chapter 6.

# Chapter 8

## Discussion

This chapter presents the discussion of this thesis research’s findings, consisting of observations made by the researcher, statements given by the participants and the results from the cognitive load questionnaires. The findings are first discussed concerning the AI assistance for both the Obslog and Debrief tasks, before examining the differences and similarities. The first two RQs about the usage and experience of the AI assistance will be answered and discussed using the researcher’s observations and participants’ statements. The findings concerning usage and experience have some overlap, because how participants used the AI assistance influenced their overall experience with it and vice versa. RQ3 regarding cognitive load will be answered and argued using the CL questionnaire results, whilst keeping the other findings in mind.

Two important themes that come up in this chapter are effectiveness and efficiency. In this thesis, effectiveness for the UX testing workflow is seen as the degree to which the objective of the task is achieved and how successful one is in producing their desired result. For obslogging this means recording relevant information that answers the client questions as thoroughly as possible. For the debrief task, this means creating a “complete” as possible and “correct” issue list (overview of key insights). How the produced obslog and issue list should look to achieve a sufficient state of effectiveness is subjective and depends on the UX researcher. Efficiency for the UX testing workflow is seen as accomplishing the task objective (to satisfaction) using the least amount of time and effort.

These variables have not been quantitatively measured, rather the discussed results have been generalised from the participants’ statements regarding their AI usage and subjective experiences. Hence instead of talking about measurable changes in effectiveness and efficiency, we look at the experienced/perceived change, which will be discussed in relation to RQ2.

### 8.1 Obslog assistance

In this section, we discuss the findings regarding the usage and experience of the AI Obslog assistance. The results concerning the cognitive load experienced during the task for both test conditions will also be examined.

### 8.1.1 Usage of AI assistance

During the experiment, participants' usage of AI assistance was investigated, which also addresses RQ1.

RQ1: *“How does AI textual data filtering assistance influence how UX researchers perform the tasks of noting down observations (and debriefing)?”*

The findings suggest that the AI Obslog assistance does not affect how the UX researchers performed the task of obslogging itself. However, participants' statements indicate that the AI assistance was used for later tasks after completing the Obslog task, such as preparing for the debrief. The usage and influence of the AI Obslog assistance will be discussed in this section for during and after obslogging.

#### During obslogging

No changes in how the UX researchers logged observations with and without AI assistance were observed. This aligns with how several participants (P1, 3-6, 8) stated that the AI assistance provided no additional value for the task at hand. P4's statement about how the AI Obslog assistance does not change the task of obslogging could explain this. The AI Obslog assistance does not tackle the challenge of the Obslog task, namely the multitasking. As identified in Chapter 3 about the UX testing workflow, the challenge of obslogging lies in multitasking and having to process and record great amounts of information simultaneously. The AI labels are supplementary to the standard Obslog labels and thus add additional information instead of reducing the amount of information participants have to process simultaneously. To avoid this, participants try to ignore the AI output and read it later. Doing so, however, defeats the purpose of displaying the AI labels in real-time and displaying the AI labels at a later moment should be considered (if they are deemed useful for later tasks). Since the AI Obslog assistance does not help with the objective of obslogging, it's difficult to evaluate it in terms of experienced effectiveness. Therefore, only experienced efficiency will be discussed for RQ2 in Section 8.1.2. Other ways AI assistance could more accurately tackle the challenges of the Obslog task will be discussed in future works (see Section 8.5).

#### After obslogging

Participants' use of the AI labels for theme formulation and quick scanning was difficult to observe because they occurred mentally without producing directly observable output. The results are therefore solely based on participants' statements, which can make it difficult to assess the potential use of AI assistance when there are ambiguous or contradicting comments. For example, P2 expressed using the AI assistance for theme formulation inspiration, but this was likely hypothetical (use of 'could' in the statements). Moreover, P5 disliked the formulation of the labels. This aligns with the concern mentioned for Gao et al. (2023)'s CollabCoder unsupervised LLM-based QC assistance about how the freedom in the AI output can result in labels that are less representative of the coder's mental model or ideas. Because of the contradicting views of the labels' formulation and the hypothetical nature of P2's comments, we conclude that the potential of the AI Obslog assistance as theme formulation is inconclusive.

P6 and P7 expressed how the AI labels helped them scan the Obslog more efficiently during Debrief preparations. This feeling of increased efficiency and factors contributing to this will be more closely examined and discussed in Section 8.1.2. Contrary to the

earlier mentioned drawback of AI labels, P7’s utilisation showcases the benefit of these labels as supplementary to the standard Obslog labels. By combining the AI output with their own inputted labels, they gained a more comprehensive understanding of the broader context, a feature they thought to be lacking for the AI Debrief assistance. The standard Obslog labels have similarities with the CollabCoder’s keyword support, which can provide context in a manner that is quicker than reading the complete source input. This shows it is possible for users to manually add support for context understanding, albeit requiring effort and time. More on context, the CollabCoder system only allowed limited user input to be incorporated into the AI output, which Gao et al. (2023) says to be crucial in guiding the AI model by supplying the nuance, context and deeper understanding that the AI may lack. The Obslog assistance attempted to incorporate more user input and intent by using the notes written by the participants. Nevertheless, those notes do not necessarily contain the desired nuance and deeper understanding or are lost in translation when used to generate the AI few-word labels. However, for future research, feeding the AI assistance the standard Obslog labels for context could help guide the AI assistance, possibly producing better results. Lastly, if the AI labels are used in P7’s manner, displaying the AI output at a later moment instead of in real-time seems to be more beneficial and suitable.

Thus, the AI Obslog assistance did not influence how the UX researchers approached the task of obslogging and the majority did not use the AI labels at all, deeming them not useful. Nonetheless, two participants used the AI labels as an addition to their standard way of working for quicker scanning of the Obslog. To better assess the impact of utilising the AI Obslog assistance in this manner, we examine UX researchers’ experience with it in the following section.

### 8.1.2 Experience with AI assistance

During the experiment sessions, the participants had a lot to say about the AI Obslog assistance, which helped answer RQ2.

RQ2: *“How do UX researchers experience the AI textual data filtering assistance during a UX testing test day?”*

Whether the UX researchers experienced the AI Obslog assistance more positively or negatively largely depended on whether they found it *distracting* or *subtle* enough to ignore when necessary, as well as whether they believed or felt like it impacted the *efficiency* of performing tasks related to the Obslog—whether it increases, decreases, or remains unchanged. Several factors contribute to the perceived efficiency impact. Those factors are illustrated in Figure 8.1 and will be discussed further in this section. Weighing the various experienced variables, participants conclude whether they find the assistance useful and whether they would want to use it in real-world projects.

#### Distractibility

The participants’ experiences of the AI labels depended partly on how distracting they found the instant AI-driven generation of labels during obslogging. UX researchers who found the AI Obslog assistance highly distracting during obslogging reported that it diverted their focus from the task, causing annoyance and disturbance and leading to a more negative view. When combined with the belief that the AI Obslog assistance provided no additional value, it resulted in a reluctance to use it in the future. For example, both P1 and P8 expressed finding the AI labels distracting and unuseful, hence they would not use



them. However, P8 did consider scanning the labels at a later moment but reconsidered due to other factors impacting experienced efficiency.

P6 found the AI Obslog assistance distracting but also useful for quickly scanning the Obslog at a later moment, leading to a more positive view and being open to future usage. However, an important consideration mentioned by P6 for whether the assistance is deemed useful is the perceived correctness of the labels. This impacts the experienced efficiency and will be further discussed later in this section. On the other hand, P7 noted that the AI Obslog assistance was subtle enough to ignore when the labels were perceived as incorrect. Combined with deeming the assistance useful, they experienced it more positively. Nonetheless, the findings suggest that the distraction regarding the AI labels is mainly due to it taking away participants' attention from the task at hand. Moreover, since the AI labels have only been used for tasks conducted after finishing obslogging, making the AI Obslog assistance not real-time is likely to solve the distraction issue without taking away possible benefits.

Thus, distractibility is a factor that contributes to participants' experience with the AI Obslog assistance. Finding the AI labels distracting is more likely to result in an overall negative experience. Displaying the AI labels after participants finished obslogging, will remove the distraction and possibly improve the experience.

## Efficiency

The participants' experiences of the AI Obslog assistance depended greatly on whether they believed it improved or could improve the efficiency of subsequent tasks utilising the Obslog. The participants with a more positive view of the AI Obslog assistance believed it improved or could improve the efficiency, whereas those with a more negative view deemed the AI Obslog assistance to decrease or not change the efficiency of any tasks related to the Obslog. Several interconnected factors that affect this experience of efficiency have been mentioned and are as follows (illustrated in Fig. 8.1):

For those who view or use the AI Obslog assistance for theme formulation: the efficiency, or in this case mental effort of coming up with a suitable term for a point or theme is only saved when the label is correct, in the sense that it summarises the corresponding 'Observation or quote' or theme well.

For those who view or use the AI Obslog assistance for a quick scan: when the UX researchers feel like the labels are correct, meaning they summarise the corresponding 'Observation or quote' well, they feel like the AI assistance *improves the scannability* of the Obslog. This means less information to process when scrolling through the Obslog, thus reducing the time and effort to do so.

Firstly, the correctness of the AI labels greatly impacts the experience of the assistance or rather the perceived ratio of correct to incorrect labels. However, the correctness of a label can be subjective since different people may find different terms more suitable to summarise a phrase. Moreover, without counting how many labels one finds correct, the amount of correct labels is based on feeling. Next, the threshold ratio of correct to incorrect labels for whether the AI assistance is usable may differ per person. Additionally, people may have different thresholds, concerning the perceived correctness ratio of AI labels, to trust the AI assistance. The general idea however is, that the more "incorrect" labels a person finds, the less trust one may have in the labels and the greater need or desire one will have to check and evaluate (all) the AI output (e.g. by completely reading the corresponding 'Observation or quote' cell), thus increasing the information to process, increasing the time

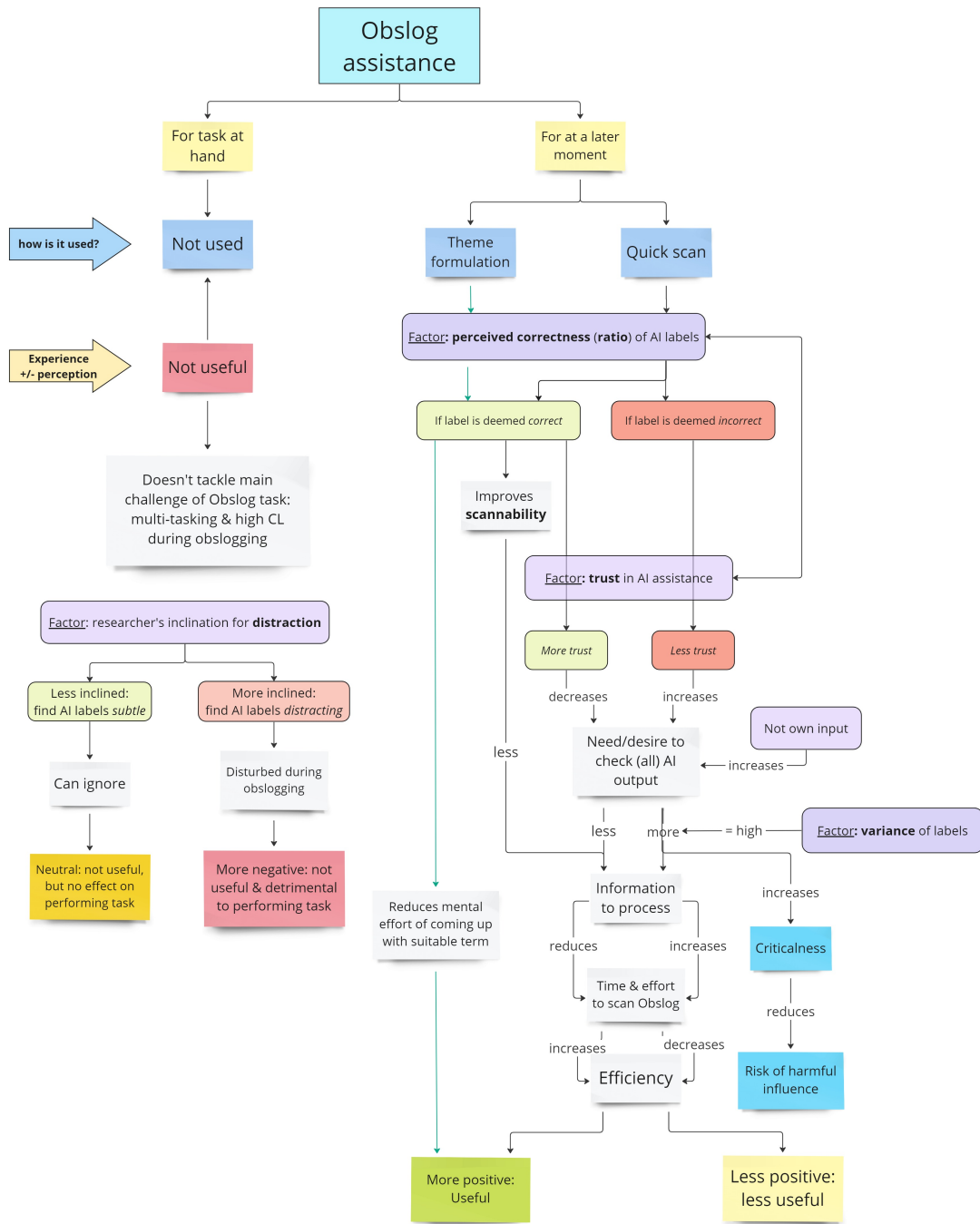


FIGURE 8.1: Influence of AI assistance on the experienced efficiency of the Obslog task

and effort needed to scan the Obslog and thus reducing the efficiency of conducting the task, or at least nullifying the potential efficiency gain that could have been produced with the AI assistance.

The perceived correctness of the AI labels strongly relates to the disadvantage of how unsupervised AI has a higher tendency to hallucinate. If the tendency for hallucination is not mitigated, users are less likely to fully trust the unsupervised AI assistance, limiting the potential efficiency gains.

The other way around, if one is already more sceptical of AI, and has less trust, one might be more critical of the AI output and perceive the number of correct labels as insufficient or feel a higher need to check the output. Moreover, the need or desire to check all the AI output might for some already be high(er) since it is not their own output. This seemed to be an important factor (in combination with distractibility) for P8 to deem the AI Obslog assistance as not useful.

Another factor for efficiency is that there is a high variance within the AI labels, meaning no terms are repeated. This also increases the amount of information to process, since in this manner the AI labels cannot function as a data filter. Rather, one has to fully read each label since each is a new term. CollabCoder’s feature of re-using labels from the code history could be applied to see whether a lower variance of labels will improve the experience of the AI Obslog assistance.

Summarising, the variables perceived correctness, trust, desire to check and evaluate AI output and variance of labels are related and influence the experienced efficiency of the task. Weighing the factors, if the eventual assessment is an increase in efficiency, the UX researcher is more likely to positively experience the AI assistance and consider it for future usage.

### 8.1.3 Cognitive Load

With the data obtained via the cognitive load questionnaires, this research aimed to answer the following research question, focusing on the Obslog task:

RQ3: *“To what extent can AI textual data filtering assistance reduce the cognitive load during a UX testing test day?”*

In short, no significant change in cognitive load during obslogging with and without AI assistance was measured. In other words, the AI textual data filtering assistance for the Obslog task did not reduce the cognitive load during a UX testing test day. This aligns with the qualitative findings of how the AI Obslog assistance was of no additional value. However, the experimental set-up did contain several limitations that could have influenced the questionnaire results, e.g. lack of anchoring for the rating questions for the two test conditions and ambiguous wording of several questions. Moreover, many participants provided the reasoning behind their ratings, mentioning other factors than the AI assistance. This raises questions on whether the difference in ratings is due to other factors than the AI Obslog assistance. If the changes in ratings between test conditions are due to other factors, then the effect of AI assistance on cognitive load is not accurately measured. This will be discussed per question type— mental effort on the Paas scale, extraneous cognitive load and germane cognitive load.

Generally, it was discovered during the experiment sessions that the participants experienced a lack of anchoring for the ratings between the two test conditions. Anchoring refers to how participants can use the rating of one question of test condition A as a reference point, an anchor, to assess what rating to give the same question for test condition B. Since the questionnaires for the tasks with and without AI were separate, participants were unable to perform anchoring and roughly estimated what they rated the previous test condition. This might have resulted in more inconsistent ratings, in the sense that participants’ ratings for two test conditions differed from their statements. To get more consistent results, the experiments should be redone using questionnaires where the participants can view their ratings for the previous test condition.

Several participants also expressed confusion on how to interpret the GCL questionnaire questions, especially what exactly is seen as an insight. So it can be debated whether everyone had the same interpretation of what the GCL questions refer to. If not, the results might be more difficult to generalise. For example, some participants think that insights only emerge during the analysis on the report-making day, whereas others already consider ideas or themes they identify during obslogging as insights. Then the ratings for improved understanding of the acquired insights are likely higher for the latter. Nevertheless, since we are interested in the difference between the ratings of the two test conditions and we employed a within-subjects study design, the ambiguous question formulation should not have a big effect on the results.

For the Paas' mental effort question, the majority of the participants' explanations behind their ratings did not mention the AI assistance. Other factors like the increased familiarity with the project topic in the second round and the talking speed of the respondent had more effect on the mental effort it took to perform the task of obslogging.

For the ECL questions, the AI labels were both mentioned to be distracting and clear in terms of functionality. The standard Obslog labels were also mentioned as too many and confusing. In addition to these comments, since the ratings concerned the whole Obslog (to be able to compare the two test conditions), it cannot be said for certain that the AI labels were the most important component for the ECL experienced for using the Obslog.

For the GCL during obslogging, many participants explained this improves naturally as the test day progresses or only during later analysis (during report making). The experimental set-up only simulated a small section of a whole test day and the test conditions were applied consecutively. Hence, it was difficult to accurately measure the effect of AI assistance on the GCL experienced during the task of obslogging.

The participants' clarifications suggest that other factors played a greater role in participants' mental effort expenditure, ECL and GCL than the AI Obslog assistance. This means that the CL questionnaire results are too unreliable to make conclusions regarding the influence of the AI Obslog assistance on reducing cognitive load during a UX testing test day. Thus, there is no evidence that the AI Obslog assistance reduces the cognitive load during a UX testing test day. However, repeating the experiments with improved questionnaires (one where participants can view their previous ratings for the other test condition and clearer question formulation) is needed for more reliable results.

## 8.2 Debrief assistance

In this section, we discuss the findings regarding the usage and experience of the AI Debrief assistance. The results concerning the cognitive load experienced during the task for both test conditions will also be examined.

### 8.2.1 Usage of AI assistance

In this section, we'll answer the following research question for the AI Debrief assistance:

RQ1: *“How does AI textual data filtering assistance influence how UX researchers perform the tasks of (noting down observations and) debriefing?”*

To do so we inspect and discuss the different views and usages/usage extents of the AI Debrief assistance. Additionally, we examine participants' evaluation of the AI output.

## Views of the AI Debrief assistance

All the participants viewed the AI Debrief assistance similarly, namely as a checklist, reminders, suggestions, a starting point, a guideline, a common thread, a summary or overview of the interview content and a partner. These descriptions all pertain to some form of assistance or additional input to help begin, continue with or complete the task of the Debrief (preparation). Or check for and fill in missing points. Even if the AI output only contains points that the participant already has, it can provide a feeling of reassurance. How UX researchers view the AI assistance reflects how they utilise it, if they do use it.

### Group 1: hypothetical or not observed usage

Half of the participants (P1-2, P5 and P7) did not visibly use the AI output to prepare for or conduct the debrief, resulting in minimal influence on their debriefing approach. At the same time, P1-2 and P5 did describe the potential usage for real projects positively. The reason their usage remained hypothetical or not visible is due to several reasons:

- Required assistance is project-dependent; they did not need the AI Debrief assistance for *this* project, already having a complete list of what they wanted to say during the debrief. Or they might have utilised it to check or add something mentally, which cannot be observed (P2, P5). Their positive outlook on future usage could be for projects or test days when they don't immediately have a complete grasp on the most important points.
- Limited experimental set-up; content for the issue list –insights, themes, etc.– organically arises after doing multiple interviews. For this experiment, after only two interviews, one would not already create an issue list (P1)
- Concerns for ethical risks; the potential harmful influence of the AI output, resulting in biased results for the debrief (P7)

For P2, it was difficult to determine their actual usage of the AI assistance. This is because of some contradicting statements:

P2: “It (the AI Debrief assistance) mostly confirmed things for me. I already had a clear view of the most important insights, so my understanding (of the insights and such) has not necessarily improved.”

This statement of P2 indicates no actual usage beyond checking and confirmation of own output. On the other hand,

P2: “I find it very effective. Usually, some points jump out for me, which I then write down (e.g. in the issue list). However, this respondent spoke very fast, so I didn't manage to do so myself. It (the AI Debrief assistance) also extracted some nuances that I did not remember myself. And it had quotes to support my point.”

This comment suggests that P2 used the AI output to add missing nuances to their debrief and utilised the AI references as evidence. Because of the contradictions, we grouped P2 as hypothetical usage. P7's comments on ethical risks will be further discussed in Section 8.2.2.

A component of the AI Debrief assistance that was only described for future usage is the AI references. These were by the majority of the participants perceived as quotes, because of incorrect presentation between quotation marks. Only a couple of participants (P6 and

P8) were able to discern that some were actually observations and that the AI references were not sourced from their own input. Although the misrepresented presentation is a sloppy mistake in the experimental setup, it also shows how difficult it can be to discern or check such information (at least in the manner the AI Debrief assistance was presented). There's a risk of being misled, meaning we should be very clear in what is provided, but also critical of what is actually given. Being critical will be further discussed in Section 8.2.2.

### **Groups 2 & 3: actual usage- various extents of influence**

The other half of the participants did visibly use the AI output to prepare for or conduct the debriefing. The consequences of using AI assistance for participants ranged from minimal influence to potentially significant effects, depending on individual approaches and degrees of reliance. P3 and P6 adopted a cautious strategy, leveraging AI output as a supplemental checklist with reminders/suggestions, to enhance the comprehensiveness of their issue lists. This method, characterized by generating one's own input first, helps mitigate potential biases and contributes to task effectiveness.

P8 initiated debriefing preparations using the AI output, risking bias towards the AI-generated points. However, their subsequent adherence to their usual approach helps mitigate potential AI bias, and enhances task efficiency, specifically during the beginning (in terms of speed and effort).

In contrast, participant P4 heavily relied on AI assistance to generate the issue list, accelerating debrief preparation but raising concerns about the task effectiveness. Although P4 did remove one AI-generated point, the difference in actions compared to the other participants, who added points and removed or modified more parts, casts doubt on the completeness and accuracy of P4's debrief. Despite expressing uncertainty about the aforementioned concerns, P4 did not verify the output, demonstrating the potential risk of harmful influence due to AI usage.

Several variables have been mentioned in regards to the usage and influence of the AI Debrief assistance –effectiveness, efficiency, ethical risk (AI bias/influence)– which will be further discussed in Section 8.2.2.

### **Evaluation of AI output**

All participants evaluated the output generated by the AI to a certain extent. Everyone expressed thoughts about whether a point is correct in the complete sense, or only partially, or only missing or having an incorrect nuance, or slightly wrong interpretation and conclusion, and also whether a point is actually relevant (enough). Such evaluation resulted in keeping, discarding, or editing specific parts of the AI output. This shows some form of criticalness (criticalness will be further discussed in Section 8.2.2) to avoid potential harmful AI influence on the debriefing. Nevertheless, P4 demonstrated that when one is being lax, they can be inclined to be less thorough in evaluating and altering the AI output to be used, even when they have some doubts about the correctness and completeness of the output. Especially for real-world UX testing projects, at the end of a long day when UX researchers might be exhausted, it is debatable whether the UX researchers will be as critical in assessing the AI output as at the beginning of a day when they're still fresh. This is similar to Gao et al. (2023)'s concerns for users' potential reliance on AI assistance when there is limited time. Thus, although all participants demonstrated criticalness to a certain degree via evaluating the AI output (before putting it to use), it should be further tested to see whether they can maintain the same or the desired level of criticalness during

a real-life project.

### 8.2.2 Experience with AI assistance

In this section, we examine participants’ experiences with the AI Debrief assistance to answer the following research question:

RQ2: *“How do UX researchers experience the AI textual data filtering assistance during a UX testing test day?”*

The more positive experience the UX researchers had of the AI Debrief assistance (compared to the Obslog assistance) seems to be largely due to whether they believed or felt like it impacts the *efficiency* or *effectiveness* of the Debrief task and preparation for it—whether it increases, decreases, or remains unchanged. Ethical risks are also taken into account by the participants. Several variables have been mentioned regarding the usage and influence of the AI Debrief assistance –effectiveness, efficiency, ethical risk (AI bias/influence)—which will be further discussed in this section, as well as the related concepts of criticalness and checking/evaluation of AI output, perceived correctness of AI assistance and trust. Similar concepts and the relation between them have been discussed for the AI Obslog assistance, but in this subsection, we’ll examine and discuss how they apply to the AI Debrief assistance.

#### Effectiveness, efficiency & ethical risk

Several connected factors have an impact on the experienced feeling of effectiveness and efficiency when executing the Debrief task. These include perceived correctness of AI output, criticalness (desire to check and evaluate the AI output) and trust. These are also connected to the ethical risks of AI, such as AI bias or harmful influence.

For the Debrief task, participants showed different levels of criticalness regarding the AI assistance, which was expressed in checking and/or evaluating the AI output or comparing the AI output with their own mental list or notes and/or the obslog. Depending on how critical the participants were, different levels of AI influence occurred. The more AI influence occurs, the higher the risk of potentially harmful AI influence/bias. At the same time, the more “good” AI influence, the higher the increase of efficiency (i.e. saving time and effort by not having to thoroughly scan the Obslog), and possibly effectiveness (by adopting points generated by AI that were originally missed). The other way around, to reduce the risk of harmful AI influence, one is likely to try to be more critical and check the AI output, reducing or not changing the overall efficiency of the task.

For participants’ criticalness, trust in the AI assistance and perceived correctness of the AI output also plays a role. As soon as participants notice that a point is not (completely) correct, their trust decreases and their need or desire to be critical and check the output increases (P8 is a good example of that). As mentioned in Chapter 2.1, current automated summarisation systems are still susceptible to errors and have a tendency or potential for hallucination (Fok et al., 2023). If these problems are not solved, users are less likely to fully trust the AI assistance, limiting the potential efficiency and effectiveness gains.

### 8.2.3 Cognitive Load

With the results acquired with the cognitive load questionnaires, this thesis aimed to answer the following research question, focusing on the Debrief task:

RQ3: *“To what extent can AI textual data filtering assistance reduce the cognitive load during a UX testing test day?”*

Shortly, no significant change in cognitive load during the Debrief task with and without AI assistance was measured. In other words, the AI textual data filtering assistance for the Debrief task did not reduce the cognitive load during a UX testing test day. This does not align with the qualitative findings of participants’ usage and experience of the AI Debrief assistance. However, the experimental setup contained several limitations that could have influenced the results, including a lack of anchoring and confusing question formulation. These limitations are the same as for the Obslog task and have already been discussed in Section 8.1.3. Moreover, many participants provided the reasoning behind their ratings, mentioning other factors than the AI assistance. This raises the same questions as for the Obslog assistance on whether the effect of AI assistance on cognitive load is accurately measured or if there is interference from other factors. This will be discussed per question type— mental effort on the Paas scale, extraneous cognitive load and germane cognitive load.

Participants’ additional clarifications for Paas’ mental effort question suggest that there was a decrease in mental effort due to AI assistance. This is contrary to the t-test results, but looking at the individual mental effort ratings we can see that P1’s rating jumps out. It was later discovered that P1 filled in a rating inconsistent with their statements due to a lack of anchoring. To ensure consistent results, the experiments should be redone using questionnaires where the participants can view their ratings for the previous test condition.

For the ECL, many participants mentioned the AI assistance as effective for debriefing, but the t-tests suggest the differences in the rating of the test conditions to be insignificant. Looking at Table 7.5, we spot a relatively big mean difference of around 2 for Q5 and Q6, supporting the additional clarification. Nonetheless, the corresponding high standard error (SE) difference, suggests the means will vary if the experiments were to be repeated with new participants. For more reliable results, experiments should be redone with more participants.

The clarification for the GCL for the Debrief task is similar to that of the Obslog task (see Section 8.1.3), concerning the naturally growing improvement as the test day progresses. Likewise, it is difficult to accurately measure the GCL with the current experimental set-up.

Thus, the quantitative results suggest that there is no significant change in cognitive load during the Debrief task with and without AI assistance. On the other hand, the qualitative findings indicate a more positive impact of the AI assistance on the debriefing. This contradiction could be because of the flaws and limitations influencing the questionnaire ratings. Repeating the experiments with more participants and improved questionnaires (one where participants can view their previous ratings for the other test condition and clearer question formulation) are needed for more reliable results.

### 8.3 Obslog & Debrief assistance

Previously, some differences between participants’ experiences with the AI assistance for the Obslog and Debrief tasks have been discussed. These include context, perceived correctness, trust and evaluation of AI output, and their effect on the perceived efficiency of the task. How these variables are related for either task is similar, but the Debrief assistance was perceived to result in more efficiency gain and thus experienced more positively. In this section, we will further examine the varying experiences in context with



the differences in the nature of the tasks and characteristics of the AI assistance provided. Furthermore, we will also debate how the integrated presentation of the AI assistance and participants' curiosity towards AI can affect people's receptiveness to using AI assistance.

Since the Obslog assistance is given in correspondence with the participants' input whilst the Debrief assistance is displayed in a different tab page in the Sheets document (assuming that participants believe the AI summary is based on their Obslog), it could be said that the AI output is presented more in context. This is consistent with P7's statements and can be said to help better understand the AI output. This also aligns with the spatial contiguity principle of the cognitive load theory, which states that related pieces of information can be better processed when presented together. Incorporating more context-providing features could improve the AI Debrief assistance. A context-providing feature could be displaying the corresponding 'Task' or 'Context' label beside the AI point. The AI references provided were intended to give context and explanation of the source of the AI output; rectifying the presentation of the references could be positive for users' experience with the AI assistance. Testing should be conducted to evaluate the possible effects of more context-providing features on the experience of the AI assistance.

Another difference between the two tasks with AI assistance is that participants had more time to process, and thus critically check and evaluate the AI output during the Debrief preparations than during obslogging. Perhaps, this could have contributed to participants being less bothered by "incorrect" points and experiencing the Debrief assistance more positively. Another explanation or factor for this could also be how many participants viewed the Debrief assistance as a checklist for missing items. Perhaps the participants were more forgiving towards AI points they disagreed with since they were already planning to only select the relevant ones. Nevertheless, more focused research has to be conducted to make any conclusion regarding factors influencing people's experience with AI assistance. For instance, having participants use the AI Debrief assistance but giving them various amounts of time to process the AI output.

The integrated presentation of the AI assistance (for both tasks) was viewed positively by the participants. The findings suggest that the integration of AI assistance into people's familiar tools, in contrast to a separate AI tool, is more efficient for the adoption of AI support. This since time and effort can be saved in regards to learning a completely new interface. Moreover, this helps lower the barriers to employing AI assistance. Additionally, curiosity also seems to contribute to people's willingness and receptiveness to try using AI tools. However, curiosity for a new tool or technology is likely to fade after more usage. Therefore, to ease the hurdles of incorporating AI assistance in (UX) workflows, integration into familiar tools can be helpful.

## 8.4 Limitations

The findings of this research have to be seen in light of several limitations. Firstly, the accuracy of how well the test day was simulated can be questioned, particularly in terms of replicating real-world pressures and mental overload experienced by UX researchers. The absence of genuine client pressure and the researcher's dual role as client may have influenced participants' behaviour and responses. This is because the debriefing procedure is heavily influenced by the (input of) the clients (e.g. if clients disagree or want to discuss something, then it'll take longer and might be more tiring. Whether the researcher needs to negate incorrect assumptions or interpretations based on partial participation, etc.), which was difficult to emulate consistently.

Similarly, participants' prior knowledge of the research objectives and their backgrounds in UX research may have biased their responses or affected their natural behaviour during the test day. Moreover, participants' limited preparations and the lack of multiple interviews could have impacted the quality and depth of insights gathered.

The small sample size raises concerns about the generalizability of the results. Although efforts were made to recruit a diverse group of participants, the limited number may not adequately account for individual differences and potential order effects. Moreover, the study's reliance on a single project's data also limits the generalisability of findings to other UX research contexts. And since this thesis is use case-based, looking at only one company's UX research workflows, it is advised to closely examine one's own typical UX testing procedures and see whether and how this research could be extrapolated to UX researchers from other teams. Additionally, the hypothetical nature of some AI influences and the lack of investigation into their potential impact on later analysis highlight areas for future research.

In conclusion, while this study provides valuable insights into the integration of AI assistance into the UX testing workflow, several limitations must be acknowledged. Addressing these limitations through larger-scale studies and real-world implementations will be essential for advancing our understanding of the potential impact of AI assistance on the UX testing workflow and its practical implications for industry professionals.

## 8.5 Future works

In the previous sections, we have discussed several flaws and limitations of the experimental setup that (may) have influenced the obtained results. To acquire more reliable and accurate results it is advised to conduct larger-scale studies in real-world settings, should one have the resources. Moreover, it would be insightful to redo the experiment of the debrief preparations with a thinking-out-loud procedure to uncover more mental processing, which cannot be visibly observed. Besides improving the experimental set-up employed for this thesis, various areas should be investigated in future works, in the journey to exploring and researching how AI assistance can (best) be incorporated into the UX (testing) workflow.

Firstly, to be able to more objectively assess the impact of AI assistance on the UX workflow (to compare with and without AI assistance), it would be imperative to define what different levels of task efficiency and effectiveness entail for the UX testing workflow (e.g. when is sufficient or desired task effectiveness achieved?) and to measure them quantitatively.

Secondly, since this research had a more explorative nature and a focus on examining the possible impact of AI assistance on the UX workflow instead of developing an ideal technology, incorporating AI assistance for UX researchers should not be limited to the approach used for this thesis. For instance, for obslogging, we used a real-time, unsupervised LLM-based approach to generate summarising labels that did not tackle the challenging aspect of the task. Rather than only trying to improve this approach and making it work by altering various characteristics, exploring other options such as automated transcription of the interviews could be profitable. Or investigating the possibilities for automated labels that could replace the 'Task' or 'Context' labels, instead of the summarising labels provided during this research. Possibly by putting (mouse) trackers in the prototype pages or other elements that can detect what elements respondents are looking at, linking this to the test script sections, and automatically filling this in for the Obslog.

Thirdly, future works should evaluate how critical UX researchers can be regarding the AI output, at the beginning vs the end of a test day. This could give more insight into the potential risk of harmful AI influence when participants are in varying conditions (i.e. mentally tired or not). Such knowledge is essential for evaluating the impact of incorporating AI assistance into any real-world workflow. Perhaps this can be investigated by conducting a between-subjects study, where each participant is given the same AI output (e.g. the AI Debrief assistance summary) to evaluate, but different participants have to perform different amounts of mentally tiring tasks (e.g. puzzles or complex arithmetics) beforehand. The amount of evaluating can be assessed based on e.g. number of points commented on, the extent of comments, etc.

Lastly, the impact of AI assistance on the UX workflow of other companies should be assessed to obtain more generalisable results. This is because the UX testing workflow used for this thesis is based on one company with certain work procedures that can differ from other UX researchers.

### **Chapter conclusion**

Thus, overall the AI Obslog assistance did not change how the UX researchers performed the task of obslogging and was experienced to be distracting and not useful for the task at hand. However, a third of the participants deemed the assistance as (potentially) useful for increasing efficiency while scrolling the Obslog at a later moment. Exploring participants' experiences with the AI assistance uncovered how perceived correctness, desire to check and evaluate the AI output, trust, and label variance are related and influence the experienced efficiency. Furthermore, no significant differences in cognitive load during obslogging with and without AI assistance have been measured.

Half the participants used the AI Debrief assistance to perform the Debrief or prepare for it. Various approaches were employed, which e.g. differed in order of when to use the AI assistance, as well as to different degrees. Correspondingly, the influence of the AI Debrief assistance on the approaches to debriefing varied. Next, the majority of the participants experienced the AI assistance positively, which is closely linked to its functionality and usage, and whether they deemed it useful. Usefulness is assessed based on perceived effectiveness and efficiency, which is affected by perceived correctness, desire to check and evaluate the AI output and trust. The differences in cognitive load during the Debrief task with and without AI were insignificant.

Lastly, comparing the AI assistance for the Obslog and Debrief tasks, the AI Debrief assistance was used by more participants and more intensively than the Obslog assistance, hence had a bigger impact on UX researchers' approach to UX testing. However, the more positive view is paired with more intense usage, as well as more ethical risks that should be considered, such as possible harmful AI influence. Features that could help improve the experience of AI assistance and ease its usage for the UX testing workflow.

Examining our findings on the AI assistance for both tasks, we suggest that adding more context-providing features, allowing users to take their time to evaluate the output, and displaying it in an integrated manner could help improve people's experience with it and ease its usage barriers.

## Chapter 9

# Conclusion

This thesis set out to investigate the potential impact of AI textual data filtering assistance on the UX testing workflow, doing so by exploring UX researchers' behaviour, experiences and cognitive load during a simulated test day. For the experiment, participants conducted the tasks of obslogging and debriefing, both with and without AI assistance.

The results of this research demonstrate that the AI assistance, while not altering UX researchers' approach to obslogging, can influence their approach to subsequent tasks related to the Obslog. Moreover, how much the AI assistance influenced how UX researchers performed the Debrief task varied, depending on the usage extent and approach. For the whole UX test day, no significant reduction in cognitive load was measured when conducting tasks with AI assistance.

Overall, the UX researchers experienced the AI assistance for the Obslog task rather negatively, whilst deeming their experience with the AI Debrief assistance more positively. This judgement was made after weighing several factors, of which (potential) usefulness is paramount. In turn, usefulness was evaluated based on perceived change in efficiency (and effectiveness). Ethical risk is also an important factor considered by some, which should be considered by all. How the UX researchers experience the AI assistance (for a specific task) influences their inclination for future usage of it.

Whether to use AI to improve one's UX workflow is a decision the company should consider thoroughly with all stakeholders. The intended functionality and usage should be considered: how will it be useful? And to what extent should we trust and check the output? If the aim of employing AI assistance is in terms of efficiency, and it is important to evaluate the AI output to avoid potentially harmful consequences, will we even be able to get a significant efficiency increase? Or are there ways to stay critical without costing too much time and effort? For the UX testing workflow, AI assistance during debrief preparations could be most useful when UX researchers are overwhelmed and mentally tired at the end of a test day, but that's also exactly when a higher ethical risk of harmful AI influence could occur.

Thus, the employed AI textual data filtering assistance could positively and negatively impact the UX testing workflow. We advise UX researchers interested in incorporating AI assistance in their workflow to be open-minded, but critical when exploring various options. Nonetheless, adding more context-providing features, allowing users to take their time to evaluate the output, and displaying it in an integrated manner could help improve people's experience with AI assistance and ease its usage barriers.

# Bibliography

- Beege, M., Schneider, S., Nebel, S., Mittangk, J., & Rey, G. D. (2017). Ageism–age coherence within learning material fosters learning. *Computers in Human Behavior*, *75*, 510–519.
- Bellman, R. (1978). *An introduction to Artificial Intelligence: can computers think?* Boyd & Fraser.
- Boyatzis, R. E. (n.d.). *Transforming qualitative information: Thematic analysis and code development*. Sage Publications.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101.
- Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 53–61.
- Chen, C.-Y., & Yen, P.-R. (2021). Learner control, segmenting, and modality effects in animated demonstrations used as the before-class instructions in the flipped classroom. *Interactive Learning Environments*, *29*(1), 44–58.
- Chen, N.-C., Kocielnik, R., Drouhard, M., Peña-Araya, V., Suh, J., Cen, K., Zheng, X., Aragon, C. R., & Peña-Araya, V. (2016). Challenges of applying machine learning to qualitative coding. *ACM SIGCHI Workshop on Human-Centered Machine Learning*.
- Engels, R. (2023, April). [https://prod.ucwe.capgemini.com/wp-content/uploads/2023/07/GENERATIVE-AI\\_-Final-Web-1-1.pdf](https://prod.ucwe.capgemini.com/wp-content/uploads/2023/07/GENERATIVE-AI_-Final-Web-1-1.pdf)
- Faber, R. (2023, June). AI: Cruciaal moment in de Geschiedenis of een hype? <https://www.ncsc.nl/actueel/weblog/weblog/2023/ai-cruciaal-moment-in-de-geschiedenis-of-een-hype>
- Fok, R., Kambhmettu, H., Soldaini, L., Bragg, J., Lo, K., Hearst, M., Head, A., & Weld, D. S. (2023). Scim: Intelligent skimming support for scientific papers. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 476–490. <https://doi.org/10.1145/3581641.3584034>
- Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., & Perrault, S. T. (2023). CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis. *CSCW '23 Companion*.
- Gebreegziabher, S. A., Zhang, Z., Tang, X., Meng, Y., Glassman, E. L., & Li, T. J.-J. (2023). PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3544548.3581352>
- Gewirtz, D. (2023, September). How does ChatGPT actually work? <https://www.zdnet.com/article/how-does-chatgpt-work/>
- Goyal, T., Li, J. J., & Durrett, G. (2023). News Summarization and Evaluation in the Era of GPT-3.

- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183, Vol. 52). Elsevier.
- Jan, J.-C., Chen, C.-M., & Huang, P.-H. (2016). Enhancement of digital reading performance by using a novel web-based collaborative reading annotation system with two quality annotation filtering mechanisms. *International Journal of Human-Computer Studies*, *86*, 81–93.
- JASP Team. (2024). JASP (Version 0.18.3)[Computer software]. <https://jasp-stats.org/>
- Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23.
- Jung, J., Seo, H., Jung, S., Chung, R., Ryu, H., & Chang, D.-S. (2023). Interactive user interface for dialogue summarization. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 934–957.
- Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior research methods*, *45*, 1058–1072.
- Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learning and Instruction*, *61*, 23–34.
- Meshkati, N. (1988). Toward developmet of a cohesive model of workload. In *Human mental load* (pp. 305–314). Elsevier Science Publishers.
- Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education*, *141*, 103618. <https://doi.org/https://doi.org/10.1016/j.compedu.2019.103618>
- OpenAI. (2017, June). Learning from human preferences. <https://openai.com/research/learning-from-human-preferences>
- OpenAI. (2022, January). Aligning language models to follow instructions. <https://openai.com/research/instruction-following>
- OpenAI. (2023a, March). GPT-4. <https://openai.com/research/gpt-4>
- OpenAI. (2023b, October). How ChatGPT and our language models are developed. <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed>
- OpenAI. (2023c, October). What is ChatGPT? <https://help.openai.com/en/articles/6783457-what-is-chatgpt>
- Örün, Ö., & Akbulut, Y. (2019). Effect of multitasking, physical environment and electroencephalography use on cognitive load and retention. *Computers in Human Behavior*, *92*, 216–229.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of educational psychology*, *84*(4), 429.
- Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science*, *32*(1/2), 1–8.
- Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. *The Cambridge handbook of multimedia learning*, *27*, 27–42.
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. [https://doi.org/10.1207/s15326985ep3801\\_8](https://doi.org/10.1207/s15326985ep3801_8)

- Paas, F., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4), 351–371. <https://doi.org/10.1007/bf02213420>
- Rietz, T., & Maedche, A. (2021). Cody: An ai-based system to semi-automate coding for qualitative research. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3411764.3445591>
- Rop, G., Schüler, A., Verkoeijen, P. P., Scheiter, K., & Gog, T. V. (2018). The effect of layout and pacing on learning from diagrams with unnecessary text. *Applied cognitive psychology*, 32(5), 610–621.
- Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.
- Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1819–1828. <https://doi.org/10.1145/1518701.1518981>
- Samoli, S., Lopez-Cobo, M., Delipetrev, B., Plumed, F., Gómez, E., & De Prato, G. (2021). AI Watch. Defining Artificial Intelligence 2.0. Towards an operational definition and taxonomy of AI for the AI landscape. URL: <https://publications.jrc.ec.europa.eu/repository/handle/JRC126426>.
- Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge University Press.
- Sweller, J., Ayres, P., Kalyuga, S., Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. *Cognitive load theory*, 71–85.
- Sweller, J., Van Merriënboer, J. J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational psychology review*, 10, 251–296.
- What is user experience? (2020, April). <https://www.productplan.com/glossary/user-experience>
- What's the difference between cx and ux? (2020, October). <https://www.qualtrics.com/au/experience-management/customer/cx-vs-ux/>
- Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 30(1), 111–120. <https://doi.org/10.1177/001872088803000110>
- Zhang, H., Liu, X., & Zhang, J. (2023). Extractive Summarization via ChatGPT for Faithful Summary Generation. *Conference on Empirical Methods in Natural Language Processing*. <https://api.semanticscholar.org/CorpusID:258048787>

# Appendices



## Appendix A

### Interview testscript

# Testscript

## AI assistentie & CL van UX researchers Experiment

Het onderzoek wordt uitgevoerd op [datum] in het UX lab van [redacted] door [UX researcher 1] en [UX researcher 2]. We voeren de test uit met [X] respondenten, bekijk de profielen in het testrooster.

### Hoofdvraag van het onderzoek

1. *To what extent can AI textual data filtering assistance reduce the cognitive load during the UX testing test day?*
  - a. *To what extent can AI textual data filtering assistance reduce the cognitive load during obslogging?*
  - b. *To what extent can AI textual data filtering assistance reduce the cognitive load during debriefing?*
2. *How do UX researchers experience the AI textual data filtering assistance?*
  - a. *During obslogging*
  - b. *During debriefing*
3. *How does AI textual data filtering assistance influence how UX researchers perform the tasks of obslogging and debriefing?*

## Introductie

ca. 5 min.

Ik heb je hier gevraagd om mij te helpen met mijn onderzoek over AI assistentie tijdens een UX test testdag. Daarom zal je tijdens dit experiment een aangepaste, kortere testdag doorlopen, waarbij je alleen de taken van obsloggen & de debrief uitvoert. Ik heb 2 video's voorbereid van al eerder uitgevoerde interviews om te kijken. Verder heb ik ook 2 varianten van een obslog (template) voorbereid. Terwijl je de "testdag" uitvoert zal ik je ook vragen om je scherm te delen (die met het obslog document).

Na elke taak zal ik je vragen een vragenlijst in te vullen, en na afloop zal ik nog wat meer vragen stellen.

Heb je daar nog vragen over?



## Voordat we aan de test beginnen, vertellen we respondenten:

- › Dat zij niet het onderwerp van de test zijn en niks fout kunnen doen. We testen een idee of een product en willen vooral weten wat wel en niet werkt.
- › Ik zal meekijken en een beetje observeren; ik kan mijn camera uit of aan doen, afhankelijk van wat jij fijner vindt.
- › Dat we de sessie gaan opnemen, maar dat de opname niet verspreid wordt en alleen voor onderzoeksdoeleinden wordt gebruikt.
- › Deel alsjeblieft je scherm waar de obslog zal staan; alles wat je niet op de opname wilt hebben, graag weghalen.

## Pre-test vragen

ca. 10 min.

### Algemene achtergrond van de respondent

Hoeveel jaar ervaring heb je met het uitvoeren van UX testen op de [REDACTED] wijze? (i.e. het gebruiken van de obslog voor notuleren en het doen van check-ins/debriefings)

Hoe oud ben je?

Hoeveel ervaring heb je met het gebruiken van AI (werk of privé)? / Heb je wel eens AI gebruikt, voor werk of privé? En wat weet je van hoe AI werkt?

Tijdens het kijken van 1 vd video's zal je AI assistentie krijgen tijdens het obsloggen en debriefen. Heb je bepaalde verwachtingen van hoe die assistentie eruit zal zien? Wat denk je dat de AI assistentie in zal houden?

## Scenario's en taken

ca. 120 min.

Terwijl de respondent de taken uitvoert, letten wij vooral op:

- › **Verwachtingen:** Werkt alle functionaliteit zoals men verwacht?
- › **Begrip:** Is alle functionaliteit en inhoud helder en begrijpelijk?
- › **Gedrag:** Waar kijkt & klikt men? Welke elementen gebruikt men wel/niet?
- › **Beleving:** Hoe ervaart en waardeert men het idee of product?



<b>Respondent</b>	<b>Obslog link met AI</b> <i>Conditie A</i>	<b>Normale obslog link</b> <i>Conditie B</i>
pilot		
R1		
R2		
R3		
R4		
R5		
R6		
R7		
R8		

<b>P#</b>	<b>1</b>	<b>2</b>
1	b + R2	a + R5
2	a + R2	b + R5
3	a + R5	b + R2
4	b + R5	a + R2

<b>P#</b>	<b>1</b>	<b>2</b>
5	b + R2	a + R5
6	a + R5	b + R2
7	a + R2	b + R5
8	b + R5	a + R2



# 1. Obsloggen

*Deel alsjeblieft je scherm waar de obslog zal staan; alles wat je niet op de opname wilt hebben, graag weghalen.*

*Zonder AI:* gewoon obsloggen zoals je normaal zou doen

*Met AI:* deze obslog die heeft wat AI assistentie ingebouwd die hier in de 'AI label' kolom zal verschijnen. De AI zal zijn best doen om een goed omvattend en beschrijvend label te genereren op basis van jouw observaties / citaten die je typt. Wat je met de AI assistentie doet en hoe is helemaal aan jou; jouw doel is gewoon nogsteeds om goed te obsloggen om later een goede debrief/analyse/rapportage te kunnen maken. En dus om o.a. de onderzoeksvragen te beantwoorden.

Als je klaar bent met obsloggen, dan kun je mij een seintje geven en dan zal ik je de vragenlijst link sturen.

Nog een laatste dingetje: misschien al voor de hand liggend, maar aub niet de video pauzeren, want dat kan je helaas ook niet doen tijdens een echte testdag ;)

**Oké, je mag beginnen!**

## Observations:

<https://docs.google.com/document/d/1GYZeeuYOLg9A4DAYw1vJ8TAov18PESWxdKPAHxeOSkQ/edit#heading=h.ng8ne0nyfg0t>

- › How does AI assistance influence **how** UX researchers perform the tasks of obslogging? + *Experience*?
- › *Zonder AI:* Hoe gebruikt de onderzoeker normaal de obslog?
- › **Let op...**
  - ◆ In hoeverre gebruikt men de kolommen? Met name context/taak/label en impact?
  - ◆ Lijkt de AI assistentie te storen?

**Vragenlijst link:** <https://forms.gle/CcXEhPWWpwHHNzCM6>

Als je wilt kun je nu even je scherm stoppen als je dat fijner vind.

Mocht er wat onduidelijk zijn in de vragenlijst, maak maar een geluidje en stel gerust vragen.



## 2. Debrief

### Debrief AI assistentie:

<https://docs.google.com/spreadsheets/d/1JAiv0SLkKN6asHZNRgei117RKpn2AOyX1WdDR5GPWUM/edit#gid=485314914>

Je hebt net een video van een interview gekeken van de UX test over de [REDACTED] site met een focus op personalisatie.  
Stel dat het interview representatief is van een hele testdag, dan zullen we zo een debrief houden waarbij je doet alsof ik de klant ben. Maar eerst geef ik je 10-15 min. om voor te bereiden, of pauze te houden; wat je normaal zou doen.

*Met AI:* in de issue list heeft de AI assistentie een bulleted samenvatting gegenereerd. Nogmaals, doe ermee wat je wilt. Je doel is om zo goed mogelijk de debrief te houden.

(Oké! Laat mij maar zien hoe je de debrief zou doen.)

### Observations:

<https://docs.google.com/document/d/1GYZeeuYOLg9A4DAYw1vJ8TAov18PESWxdKPAHxeOSkQ/edit#heading=h.ng8ne0nyfg0t>

- › How does AI assistance influence **how** UX researchers perform the tasks of (preparing for the) debriefing? + *Experience?*
- › *Zonder AI:* Hoe gebruikt de onderzoeker normaal de issue list?
  - ◆ Waarmee beginnen ze? Hoe structureren ze hun punten/inzichten, als ze wat opschrijven?
- › **Let op...**
  - ◆ In hoeverre gebruikt men de AI output (punten & refs)? En hoe?
  - ◆ In hoeverre passen ze de AI output aan?
  - ◆ In hoeverre voegen ze hun eigen toevoeging toe?
  - ◆ Lijkt de AI assistentie te storen? Te (veel) beïnvloeden?

Vragenlijst link: <https://forms.gle/mCpCaT2VaEo4siBp8>



# Evaluatie

ca. 30 min.

## Ervaring

Je hebt zojuist 2x geobslogged en een debrief gedaan voor 2 interviews. Bij eentje kreeg je AI assistentie.

Hoe was je ervaring? / Wat is je indruk? / Hoe was dat?

Hoe ging het obsloggen / debriefen (met AI assistentie)?

- En in vergelijking zonder AI assistentie?
- Wat vond je van de **inhoud** van de labels/samenvatting?
- Wat vond je van hoe de assistentie werd (**weer**)gegeven? Hoe het opgesteld is? (i.e. per regel, in z'n eigen kolom, in een aantal woorden)

In hoeverre heeft de AI assistentie je geholpen met het obsloggen / debriefen?

Hoe zou je nu verder gaan (om te analyseren), op de analysedag / einde vd testdag?

\* *Vraag naar **observaties** die ik al heb gemaakt* [verificatie & uitleg]

**Hoe? Waarom (zo)?**

## Vragenlijst antwoorden

Redenen voor vragenlijst antwoorden

Ik zie dat je voor <Taak> & <test conditie> een ... hebt gegeven. Waarom gaf je een <rating> voor deze vraag?

^^ sws voor Paas vraag; VERGELIJK TEST CONDITIES voor beide taken

## +/- & Verbeter suggesties

Aan het begin van het experiment vertelde je over je verwachtingen van de AI assistentie. In hoeverre kwam wat je net kreeg overeen met je verwachtingen?

In hoeverre is AI assistentie tijdens een testdag, zoals je die nu hebt gezien, voor jou interessant? Waarom (niet)?

Wat zou er veranderd moeten worden aan de AI assistent om het voor jou (nog) interessanter te maken? / Zijn er dingen die je nog mist?

Wat zou er volgens jou beter kunnen? Wat vond je goed/slecht? Waar liep je tegenaan?

---

**WAAROM? HOEZO? IN HOEVERRE?**





## Appendix B

# Cognitive Load Questionnaire

The following questionnaire is the obslog variant; the one for the debrief task is almost identical, except for some formulations adapted to the debrief task. Moreover, the client's name is redacted for GDPR reasons.

# Vragenlijst - Obslog

Vul alsjeblieft de volgende vragen in. Er zijn geen juiste antwoorden, dus geef een cijfer naar wat jij voelt dat het beste jouw ervaring representeert.

Als de vragen niet duidelijk zijn, aarzel niet om de onderzoeker voor verdere toelichting te vragen.

## \* Verplichte vraag

---

1. Dit was test conditie \*

Mocht je het niet meer weten, vraag dit gerust na bij de onderzoeker

*Markeer slechts één ovaal.*

A

B

2. Ik vond het uitvoeren van deze taak, het obsloggen, mentaal... \*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9

Hele           Ontzettend inspannend

3. Als je verdere toelichting wilt geven op de vorige vraag:

---

4. De onderwerpen die opkwamen tijdens het interview van het XXXXXXXXXX project waren... \*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

Hele            Ontzettend complex

5. Als je verdere toelichting wilt geven op de vorige vraag:

---

6. De termen die gebruikt werden tijdens het interview van het [REDACTED] project waren...

\*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

---

Hele            Ontzettend complex

---

7. Als je verdere toelichting wilt geven op de vorige vraag:

---

8. De structuur van de opgedane informatie tijdens het interview vond ik... \*

De structuur refereert naar o.a. de lineariteit van het interview, bijvoorbeeld denk aan of het gesprek een stapsgewijze prototype volgt, of dat de besproken taken/onderwerpen in een willekeuriger volgorde gebeurden; hoeveelheid belangrijke bijvangst; of er ook nog een beetje behoefte onderzoek aan te pas kwam, etc.

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

---

Hele            Ontzettend complex

---

9. Als je verdere toelichting wilt geven op de vorige vraag:

---

10. De rollen van de kolommen in de obslog waren... \*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

Hele           Ontzettend duidelijk

11. Als je verdere toelichting wilt geven op de vorige vraag:

\_\_\_\_\_

12. De obslog kolommen, in relatie tot het helpen met observaties notuleren, waren... \*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

Hele           Ontzettend effectief

13. Als je verdere toelichting wilt geven op de vorige vraag:

\_\_\_\_\_

14. Het doen van deze taak, het obsloggen, heeft mijn begrip van het behandelde project onderwerp... \*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

Hele           Ontzettend verbeterd

15. Als je verdere toelichting wilt geven op de vorige vraag:

\_\_\_\_\_

16. Het doen van deze taak, het obsloggen, heeft mijn begrip van de verkregen inzichten (opgedaan tijdens het interview)... \*

Met inzicht bedoelen we observaties en quotes of verdere interpretaties daarvan die helpen om de onderzoeksvragen te beantwoorden en/of op een andere manier mogelijk meegenomen zullen worden in de analyse & het rapport

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

Hele           Ontzettend verbeterd

17. Als je verdere toelichting wilt geven op de vorige vraag:

---

18. Het doen van deze taak, het obsloggen, heeft mijn begrip van de verbanden tussen de verkregen inzichten en de bijbehorende thema's... \*

*Markeer slechts één ovaal.*

1 2 3 4 5 6 7 8 9 10

Hele           Ontzettend verbeterd

19. Als je verdere toelichting wilt geven op de vorige vraag:

---

---

Deze content is niet gemaakt of goedgekeurd door Google.

Google Formulier

# Appendix C

## Transcript translated excerpts for Results

### C.1 Obslog

#### C.1.1 Usage

##### Theme formulation

P2: “Want hij helpt misschien weer net met op een formulering komen waar ik zelf niet op was gekomen [...] Heb je een zin en een omschrijving in je hoofd. Dan denk ik oh maar dat is wel een fijn woord om te gebruiken om dat soort van samen te vatten. Het is een partner die samen met je meedenkt. En vaak dan zegt een collega een woord en dan zegt oh ja oké laten we die soort van gebruiken. Dat is een beetje hoe dit dan voor mij ook kan werken.”

P2: “The AI assistance could work like a partner who thinks along with you; like a colleague who uses a specific word of which you think, yes let’s use that.”

##### Quick scan

P6: (About the AI ‘Obslog’ assistance) “In 1 woord te kunnen zien wat de strekking van haar verhaal is. [...] Dan lees je eerst het AI label, en als ik dan denk daar wil ik meer over lezen dan lees ik dit (de observatie/quote).”

P6: (About the AI Obslog assistance) “In one word, it shows the main message of the observation/quote, of her story. [...] Then you can first read the AI label, and if I think that I want to know more about it, then I’ll read this (the observation/quote).”

#### C.1.2 Experience

##### Distractiveness

P8: “Ik ben sowieso best wel snel afgeleid en als er iets anders dan op mijn scherm gebeurt of een pop upje, alles staat bij mij uit tijdens interviews. Dus als ik iets ineens woorden zie bewegen. Ik verwonder er maar over. Ik dacht oh ja, ik snap het wel. Of soms dacht ik nee, dit is echt totaal niet wat het is. En

dan was mijn gedachte dus bij dat label en niet meer bij het interview. En nou ja, heb ik misschien een quote gemist of een op een observatie gemist? Dus op dat moment, zeg maar. Tijdens het typen vond ik het vervelend”

P8: “I am easily distracted if something happens on my screen, so I have everything (like pop-ups) turned off. If I suddenly see some words move then I immediately think about it, I thought ‘oh yes, I agree or no that’s definitely not right’. But then my mind was on the AI labels and not on the interview/task at hand, and then I might have missed a relevant quote or observation. So at that moment, during typing (logging observations) I found it very annoying and disturbing.”

P6: “Ik ga het niet lezen terwijl ik aan het schrijven ben.” (it would take away one’s attention from the task)

## Efficiency

P6: “Sometimes he (AI Obslog assistance) summarises it (P6’s observations/quotes) really well [...]. That he writes it in two words is very handy. I think when I can trust that the AI assistance can do this well, it could be handy.”

“If the labels are mostly pretty accurate, then I think I could really use them. Because then it will be very easy to scan (the Obslog). Then I don’t have to scan all my observations to derive my conclusions.”

P6: “Soms dan vat ie het wel heel goed samen, bijvoorbeeld op een gegeven moment deed hij best wel dat ik zei van, iets met inclusiviteit was dat dat hij dat gewoon in twee woorden opschrijft, dan is het allemaal handig. Ik denk dat als ik de AI assistentie kan vertrouwen om dit goed te doen, dan zou het heel handig zijn.”

“Als de labels allemaal best wel accuraat zijn, dan zou ik hem denk ik wel echt gebruiken. Want dan is het gewoon heel makkelijk om te scannen en niet al mijn observaties te hoeven lezen, en dan gewoon daaruit een conclusie te trekken eigenlijk. ”

P3: “The AI labels should all be correct to be able to use them. Because else you will have to check every time, is the label even correct? And then you’ll be doing double (checking) work, and you’ll maybe be better off just checking out the observations.”

P3: “ Dus kijk, het zou wel allemaal goed moeten zijn om ze te kunnen gebruiken, denk ik. Want anders ga je, moet je elke keer toch checken, klopt het label überhaupt? Bij wat er getypt is. En dan ben je eigenlijk dubbel werk aan doen, dan kan je misschien beter gewoon kijken naar de observatie alleen.”

P8: “Because I did not type them (AI labels) myself, it makes (processing) it a little different. Then I have to look at it more thoroughly to see what it is about.

P8: “Alleen ik heb ze zelf niet getypt en dat maakt het net toch anders. Dan moet ik even twee keer kijken van oh ja, waar zat dit ook alweer?”

P8: “(Talking about the many different labels) Then I would have to read the

whole column to understand and process the all. [...] Yes, looking at it (the AI Obslog assistance) again, the labels don't have any additional value for me at this moment. This is because there are so many different labels, still resulting in a hundred separate items."

P8: "(Talking about the many different labels) dan zou ik bijna die hele kolom moeten gaan lezen omdat weer te snappen"

"Ja, als ik er nu zo nog een keer naar kijk, dan heeft de label voor mij. Op dit moment is geen toegevoegde waarde omdat het zoveel verschillende zijn, dus dan heb ik alsnog honderd losse items."

## C.2 Debrief

### C.2.1 View of AI debrief assistance

#### Checklist

P5: "Looking at it now (the AI Debrief assistance) I can imagine that you can nicely compare it with your own points. Like, have I seen everything? Is it complete? Because there will be a moment (a test day) where you'll forget a certain aspect or point."

P5: "Ik kan me voorstellen, als ik dit zo doorneem, [...] maar dit kun je er dan lekker langs houden (with one's own points). Van heb ik alles gezien? Ben ik volledig? Want het komt er gewoon voor dat je op een gegeven moment even nog een bepaald aspectje vergeet."

P6: "I thought it was pretty cool that he (the AI Debrief assistance) could give suggestions in the issue list. Because sometimes I'm quite overwhelmed with all the obtained information, and then it is quite nice if he provides suggestions."

P6: "En ik vond het best wel vet dat hij in de issue list suggesties kan geven. Want soms dan ben ik best wel een beetje van, wow, een beetje overwhelmed van alle informatie die ik heb gekregen en dan is het wel fijn als hij suggesties geeft."

P7: "I thought the first few points he (the AI Debrief assistance) gave were very useful. They were correct and accurate, and for me they were a nice reminder."

P7: "[...] de eerste paar punten die hij gaf, die vond ik heel nuttig. Die waren correct en die waren accuraat en die waren voor mij een nice geheugensteuntje." "Alleen ik vond het wel fijn om even snel een terugblik te hebben naar wat deze respondent nou vond."

#### Starting point

P2: "Gave direct, relevant input to structure the Debrief."

P2: "Heel prettig! Geeft direct relevante input om de debrief mee te kunnen structureren."

#### Summary, overview

P7: "Nice and clear list of insights that do a pretty good job at summarising what has been said"



P7: “Ook vooral omdat dit gewoon een duidelijke lijst is van dingen die onder elkaar staan.[...] Goede samenvatting, en goede punten genoemd.”

P8: “All in all it’s a good summary. It helps you zoom out and write down overarching insights.”

P8: “Dus over het alles heen genomen is het een mooie samenvatting. Het helpt je eigenlijk uitzoemen en overkoepelende inzichten te schrijven.”

P7: “I thought it was nice to have a quick review, an overview of what the respondent had said and thought.”

P7: “Alleen ik vond het wel fijn om even snel een terugblik te hebben naar wat deze respondent nou vond.”

#### Common thread, guideline

P6: “Er stonden wel een aantal dingen in... dat ik dacht van dat dat wel helpt... qua zeg maar rode lijn.”

“Actually, it is very nice that there is already some kind of guideline here. It would be especially handy if you would get this list for each respondent, each interview! Then you could scan on the report-making day and check if there’s a common thread or any outliers within the respondents.”

P2: “Eigenlijk is het heel fijn dat hier een beetje al een soort van leidraad staat. Vooral als dit lijstje voor elke respondent, elk interview wordt gemaakt is dat ontzettend handig! Dan kan je op de analysedag scannen en kijken of er een rode draad is of juist uitschieters bij een van de respondenten.”

#### Partner

“It could be nice for when you’re checking the points written in the issue list. It’s kind of like there is a second UX researcher who noted down their own findings. Because sometimes you miss some points that could be relevant, and then I’ll think like oh yes, that’s a good one (referring to a point mentioned by their partner). It’s as if you’re not on your own, but that there is someone else who also watched the interview and can provide input for the Debrief.”

P3: “Ja, misschien mooi voor als je er van checkt dan. Van, oh ja, mijn partner of weet ik veel wat die... Want je mist soms gewoon dingetjes die wel en toch wel goed zijn. En dan denk je, oh ja. Het is een beetje alsof er toch nog een tweede onderzoeker is die zijn of haar eigen findings heeft genoteerd. Dat je hier niet in je eentje voor staat, maar dat het toch nog iemand is die ook heeft meegekeken.”

### C.2.2 Usage of AI debrief assistance

#### Group 1: hypothetical or non-usage

P1: “I can imagine that it is interesting; not to just copy-paste, but to check if the AI mentions something interesting that I missed, like oeh that’s interesting, I’ll include that (in the issue list and debrief)”.

“I would be happy to read it (the AI list) before the debrief starts.”

P1: “Ja, dan kan ik me voorstellen dat dit wel interessant is. Niet om dit klakloos over te nemen, maar om gewoon even te kijken van... Hé, heeft AI nog iets genoemd wat ik niet heb meegenomen, wat misschien wel interessant is?”  
“En dan zou ik er blij mee zijn dat ik dat even kan doorlezen.”

P5: “Ik kan me voorstellen, als ik dit zo doorneem, dat het best wel matcht met de punten die... Ja, er zitten een paar hele gekken tussen, maar dit kun je er dan lekker langs houden. Van heb ik alles gezien? Ben ik volledig? Want het komt er gewoon voor dat je op een gegeven moment even nog een bepaald aspectje vergeet.”

“Dus opnieuw, ik zou dit als een soort eerste aanzet gebruiken. Dan denk ik nou, een stuk of vijf, zes van deze punten kan ik min of meer overnemen.”

P2: “It (the AI Debrief assistance) mostly confirmed things for me. I already had a clear view of the most important insights, so my understanding (of the insights and such) has not necessarily improved.”

P2: “Het heeft met name voor mij bevestigd. Ik had zelf al een helder beeld van de belangrijkste inzichten, dus begrip is niet zozeer verbeterd. Maar wel heel prettig op een rij.”

P2: “I find it very effective. Usually, some points jump out for me, which I then write down (e.g. in the issue list). However, this respondent spoke very fast, so I didn’t manage to do so myself. It (the AI Debrief assistance) also extracted some nuances that I did not remember myself. And it had quotes to support my point.”

P2: “Heel effectief. Normaal gesproken springen er een aantal zaken voor mij uit en die zet ik dan al op een rijtje. Deze respondent sprak ontzettend snel. En ik redde het daarom niet om zelf te doen. Ook haalde het wat nuances naar boven die ik zelf niet had onthouden én had het quotes om mijn punt te ondersteunen.”

## **Group 2: addition to standard approach**

P3: “I generated my own insights and now I’ll compare it with what the AI produced.”

P3: “Ik heb mijn eigen inzichten gedaan en nu zal ik het vergelijken met wat AI had gedaan.”

AI Debrief output: 9. Filtering of information and inspiration are both desired.  
P3 issue list: Filtering of information and inspiration are both desired.

AI Debrief output: 9. Filteren van informatie en inspiratie zijn beide gewenst.  
P3 issue list: Filteren van informatie en inspiratie zijn beide gewenst.

AI Debrief output: 4. There is a need for interaction and more visual appeal.  
P3 issue list: There is a need for interaction.

AI Debrief output: 4. Er is behoefte aan interactie en meer visuele aantrekkelijkheid.

P3 issue list: Er is behoefte aan interactie

P3: “Need more interaction. I already had the visual appeal, but I thought that point was very good. ”

P3: “Meer behoefte aan interactie. Die visuele aantrekkelijkheid had ik al. Maar die vond ik ook nog een hele goeie.”

AI Debrief output: 10. The reading list is seldomly looked at.

P3 issue list: The reading list won't be used.

AI Debrief output: 10. De leeslijst wordt zelden teruggekeken.

P3 issue list: De leeslijst wordt niet gebruikt.

P6: “I didn't literally copy-paste it. I just scanned the AI points whilst thinking ‘what is this about again?’ And then I supplemented what I had already written down (in the issue list) using the AI points, filling in any gaps.”

P6: “Ik zeg maar ik heb ze niet letterlijk overgenomen. Het was gewoon meer dat ik ging kijken van... wat staat hier ook alweer? Waar ging het over? Ja, ik ben echt heel doorheen gaan scannen eigenlijk. En toen heb ik het een beetje aangevuld wat ik hier al had staan.”

AI Debrief output: 6. An account has little additional value.

P6 issue list: (Under ‘make account’ header) making an account goes well, but has little additional value.

AI Debrief output: 6. Een account heeft weinig toegevoegde waarde.

P6 issue list: (Under ‘account aanmaken’ header) account aanmaken gaat goed, maar weinig toegevoegde waarde

P6: “For example, ‘an account has little additional value’ is a really good one; that is indeed important, so I added it (to their issue list)”.

P6: “bijvoorbeeld ‘een account heeft weinig toegevoegde waarde’ dat is ook echt een goeie; dat is wel inderdaad wel een belangrijke dus die heb ik daarmee ook aangevuld”

P8: “In the end, I created my issue list with the help of the AI Debrief assistance. I examined the AI points, thinking about what each point is about, what is missing, etc. Next, using the AI points I filled in my issue list, but I did go through the Obslog again to check for any other, missing information to make the issue list more complete. Sometimes you have time for that, sometimes you don't (to check the Obslog). In this case, I did have the time. [...] I did (often) miss what test script sections corresponded with the AI points, so I eventually went ahead and wrote those topics down myself.”

P8: “Maar ik ben dus uiteindelijk toch wel zelf mijn lijstje gaan schrijven met behulp van wat er boven stond (referring to the AI points).”

“Ik ging eigenlijk kijken van oké, wat staat hier bovenin? Oké, dit gaat dus over verschillende platformen en het zoekgedrag dan over wat ze waarderen dan behoeftes en navigatie. Ik dacht van oké, ik mis hier de plekken waarop het plaatsvindt.”

“met behulp van wat er boven stond heb ik dit een beetje aangevuld en daarna ben ik wel nog door de opslag heen gegaan om te checken. Wat dit Ja dekt dit nu alles of niet? En het is nu een interview, dus dat is heel makkelijk te doen. Soms heb je daar tijd voor, soms niet. In dit geval wel”

“In ieder geval die de plekken waarop het plaatsvond en dus de onderdelen van het script die die miste. En daarom ben ik uiteindelijk dus die verschillende onderwerpen gaan opschrijven.”

AI Debrief output: 1. The respondent uses multiple apps and mailing lists for pregnancy information.

2. The respondent looks for trustworthy sources and reads multiple articles for information.

P8 issue list: Platform | Apps/ mailing lists/newsletters; use multiple sources (added later, after scanning Obslog: and looks for familiar sources) but does not want it to become too much (information)

AI Debrief output: 1. De respondent gebruikt meerdere apps en mailings voor zwangerschapsinformatie;

2. De respondent zoekt betrouwbare bronnen en leest meerdere artikelen voor informatie;

P8 issue list: platform | apps/mailings/nieuwsbrieven; gebruiken meerdere bronnen (added later, after scanning Obslog: en zoeken naar bekende afzenders)-maar willen niet dat het teveel wordt

AI Debrief output: 3. The respondent appreciates visual content, like videos and visualisations.

P8 issue list: growth calendar | appreciates visual content, videos and visualisations

AI Debrief output: 3. De respondent waardeert visuele content, zoals video's en visualisaties; P8 issue list: groeikalender | waarderen visuele content, videos/visualisaties

AI Debrief output: 4. The respondent has a need for clear and structured navigation, and summaries of the articles

P8 issue list: growth calendar | misses navigation

AI Debrief output: 4. De respondent heeft behoefte aan een overzichtelijke navigatie en samenvattingen van artikelen;

P8 issue list: groeikalender | missen navigatie

AI Debrief output: 5. The respondent finds transparency very important when making an account and sharing personal data.

P8 issue list: want more information on...

changed to: making an account should be more transparent.

AI Debrief output: 5. De respondent vindt transparantie belangrijk bij het aanmaken van een account en het delen van persoonlijke gegevens;

P8 issue list: willen meer info over...

changed to: transparanter zijn in aangeven van account.

P8: (First comment) “I thought, ‘Oh, I’ll go and copy-paste, and select points”  
(At a later moment) “There were a couple points that gave more emphasis on certain parts that I did not find as important [...] So the AI list contained some stuff what made it that I couldn’t copy the points completely as they were.

P8: “Ik dacht al van oh ik ga dingen copy paste er uithalen”

“Er waren een aantal in waarvan ik dacht ja, dit heeft meer nadruk gekregen

dan ik vond dat het belangrijk was. [...] En daar zitten dus dingen in waarvan ik zeg ik kan niet één op één deze zinnen kopiëren.

### C.2.3 Evaluation of AI output

P4: (At first glance) “looks to be very similar to what I wrote down the 1st round/ interview”

P4: (At first glance) “Lijkt erg vergelijkbaar met wat ik de 1e ronde heb opgeschreven (in their paper Debrief list)”

P2: “I did still have to mentally consider, think over the motivation -the why-behind the AI points. Now it (the AI points) says what happens, but not why it happens [...] I look at each point, and think what can I keep, change and discard?”

P2: “Ik moest alleen mentaal nog goed nadenken over de motivatie - de waarom - achter de punten die eruit gefilterd waren. Nu staat er wat er gebeurt, maar niet waarom/de aanleiding/het gevolg van hetgeen wat zich voordoet.”  
Bij elk punt kijk ik, Keep/change/discard.

P6: “I still wanted to briefly check it myself, so then I scanned and reviewed the AI points.”

P6: “Maar ik wou het toch zelf nog even checken. Dus toen ben ik wel hier (the AI points) doorheen gelopen.”

P5: (Referring to the AI point: There is too much content, resulting in an overload.) “I don’t think the respondent actually said there was too much content. They stated that there was too much information on one screen. I think that is different from there being too much content.”

P5: “–Er is te veel content waardoor het een blur wordt.– Zij heeft volgens mij niet gezegd dat dat te veel content is. Zij heeft gezegd dat ze te veel op één scherm zag. Dat vind ik iets anders dan dat er te veel content zou zijn.”

P5: “(Referring to the AI point: Account had little additional value) I think this is definitely a correct and valid conclusion. So that point, is correct. The respondent gave various reasons for it.”

P5: “–Account heeft weinig toegevoegd–, waarvan ik denk dat dat helemaal een terechte conclusie is. Dus die is, die klopt. Zij gaf daar meerdere redenen voor”

P5: “(Referring to the AI point: an app is preferred over a website) I did not hear the respondent say this. She said something about it, but in a totally different manner. Hence, I finds this a completely wrong insight.”

P5: “–een app heeft de voorkeur boven een website– ik heb dat niet bij haar teruggehoord. Ze heeft daar iets over gezegd, maar op een totaal andere manier. Dus dit vind ik een totaal verkeerd inzicht.”

P7: “(Referring to the AI point: 8. The respondent looks for the source and references to assess the trustworthiness of the magazine.) I don’t think this matches with what the respondent said. I don’t know for sure whether the AI

interpreted this correctly. I think this refers to the fact that it belongs to a bigger organisation. And whether that influences the magazine's trustworthiness or not, ehm..."

P7: "–Punt 8: De respondent zoekt naar bronvermelding en verwijzingen naar het magazine voor betrouwbaarheid.– Ik weet niet in hoeverre dit overeenkomt met wat ze heeft gezegd. Ik weet niet zeker of de AI dit goed heeft geïnterpreteerd. Volgens mij ging dit over het feit dat het onder een grotere moederorganisatie viel. En of dat kan invloed was op de betrouwbaarheid of niet. Ehm..."

P8: "There were several (AI points) of which I thought that some parts had gotten more emphasis than I thought was necessary. For example, 'the respondent looks for trustworthy sources and reads multiple articles to check for correctness', is true. However, that the respondent is constantly looking for trustworthiness is not true. It is recognisable, not trustworthy. So, there are a couple of things which make it that I cannot outright copy the AI points word for word.

P8: "Er waren een aantal in waarvan ik dacht ja, dit heeft meer nadruk gekregen dan ik vond dat het belangrijk was. Dus de respondent zoekt betrouwbare bronnen en leest en leest meerdere artikelen dat ze meerdere artikelen leest en dat ze daarmee verifieert of klopt, dat klopt. Maar dat zij continu op zoek is naar dat betrouwbaarheid, dat is niet zo. Het is herkenbaar, niet betrouwbaar. En daar zitten dus dingen in waarvan ik zeg ik kan niet één op één deze zinnen kopiëren."

P7: "I'm not sure whether this was the crux of the interview, of what the respondent said. I think it was more of a, 'Oh, this is not necessary so I won't tick the checkbox.'"

P7: "Ik weet niet zeker of dit echt een kernpunt was wat naar voren kwam uit het (interview). Volgens mij was het meer van, oh het is niet verplicht dus ik vink het ook niet aan."

P5: "But (AI) point three is spot on. Yes, clearer headers are needed for better navigation. That is... I would maybe not use the term navigation, because that suggests going to different pages, but this was on the same page. I would adjust that. For the rest, this topic resonates well with what the respondent said."

P5: "Maar bijvoorbeeld die nummer drie, die is spot on. Ja, duidelijke kopjes gewend voor betere navigatie. Dat is... Ik zou dan misschien weer niet de term navigatie gebruiken, want dat suggereert dat je naar andere pagina's gaat. Maar het was binnen de pagina. Dat zou ik in ieder geval aanpassen. Dat vond ik een topic... Die bij haar best sterk was inderdaad"

P3: "(Referring to the AI point: the app is preferred over a website) I found this a pretty good one. Although, it was mentioned more in passing."

P3: "En oh ja, de app heeft een voorkeur boven een website. Dat vond ik ook wel een goeie. Maar dat was meer een soort terloops genoemd.

P7: "(Referring to the AI point: The respondent appreciates visual content,

like videos and visualisations.) I already listed this point. I think this point refers to the respondent's need for quick and clear information.”

P7: “De respondent waardeert visuele content, zoals video's en visualisaties. Dat noemde ik ook al. En dat slaat denk ik ook weer op de behoefte die zij heeft aan snelle duidelijke informatie.”

P6: “I felt like it did not always give the most important insights [...] It provides a couple of important insights and insights that are less important. Then I select which ones I find important enough to write down in the issue list.”

P6: “ik had nog niet helemaal het gevoel dat het echt de meest belangrijke inzichten noemt. [...] Het noemt dan een aantal inzichten die wel belangrijk zijn en een aantal inzichten die minder belangrijk zijn. En daar kan ik dan weer uithalen welke ik dan belangrijk vind om hieronder op te schrijven.”

#### C.2.4 AI references

P2: “The addition of quotes is very nice, because the client often looks for evidence of what we claim as insights, and it helps illustrate the points. Then clients cannot say that it is merely the interpretation of the UX researcher[...] It would be ideal if you have multiple quotes for a point/ insight. Then you don't have to go through the Obslog again.”

P2: “De aanvulling van de quotes is ook prettig, want de klant zoekt vaak onderbouwing/dan spreekt het tot de verbeelding. Kan er nooit geclaimd worden dat het interpretatie van de onderzoeker is.”

“Als je meerdere quotes zou kunnen hebben die bij dat inzicht horen. Dat zou natuurlijk ideaal zijn. Want dan hoeft je zelf niet nog een keer door de obslog.”

“I quite quickly got the idea that they (the AI references) were not only quotes that I wrote down (created by the participant typing it in quotation marks in the Obslog) but that it also showed what the observations I wrote down as quotes.”

P8: “En daar had ik al vrij snel het idee dat het niet alleen maar quotes waren, als in niet alleen maar cellen die ik met aanhalingstekens heb, maar dat het ook quotes zijn van delen die ik gewoon als observatie heb opgeschreven.”

#### C.2.5 Ethical risks

P1: “I spot some risks if we just give (the client) this top 10 (the AI Debrief assistance).”

P1: “Ik zie wat risico in als we alleen deze top 10 [AI debrief assistance] geven”

P6: “I looked a little bit at this (the AI points), but I didn't want to get too influenced by it. Hence, I first wrote down points in the issue list based on what I remembered myself.”

P6: “ Ik heb hier een beetje naar gekeken... maar ik wou me nog niet te veel door laten leiden (door AI assistentie). Dus ik heb eerst echt ingevuld op basis van wat ik zelf herinner.”

P7: “I found the first few AI points very useful; they were correct and accurate and served as a nice reminder. However, the subsequent points contained aspects that the AI interpreted incorrectly. If I so happen to be a little lax as a researcher, taking the AI point as the truth, then I’ll have incorrect results. And that can be dangerous. So yes, then I would rather not use it so you don’t have that danger.”

P7: “Of ja, het is een beetje het ding, want de eerste paar punten die hij gaf, die vond ik heel nuttig. Die waren correct en die waren accuraat en die waren voor mij een nice geheugensteuntje. Alleen daarna zag ik dingen die hij verkeerd had geïnterpreteerd. Ja, als ik dan even lax ben als onderzoeker en ik neem dat maar aan voor waarheid, dan heb ik resultaten die niet meer kloppen. En dat kan gevaarlijk zijn. Ja, dus dan heb je liever het gewoon niet te gebruiken zodat je niet dat gevaar hebt dan.”

## C.2.6 Experience of AI debrief assistance

### General positive responses

P1: “I would be happy to read it before the debrief starts. [...] Het is fijn dat het er is.”

P2: “Maar indrukwekkend hoor dit. Ik vind dit echt top, de AI assistentie”  
"Vind het echt heel cool om te zien”

P2: “Heel prettig! Geeft direct relevante input om de debrief mee te kunnen structureren.”

P3: “Ik ben echt verrast door de conclusies. Want die kwamen wel redelijk overheen met wat er over was/wat ik had”

P5: “Ik vind het wel interessant. Ik vind dit wel van een heel ander niveau dan wat er in dat ene kolommetje staat (referring to AI Obslog assistance)”

P8: “Toen ging ik daarheen en toen las ik de eerste regels. En toen hier was dit was mijn eerste reactie, was echt wow, Waar? Hoe? Waar komt dit vandaan? Maar wel heel vet.”

### Effectiveness & Efficiency

P2: “The AI input really helped with getting the most important points from the interview, which saves time and effort.”

P2: “De AI input hielp al heel erg met de belangrijkste punten uit het gesprek halen. Dus dat scheelt weer”

P5: “I don’t think the AI made it more effective, but perhaps it did make it more efficient. For me, effectiveness refers to achieving what you want, and for efficiency it is about how much effort it takes you. This AI assistance reduces the amount of effort needed.”

Although P5 stated that the AI assistance did not help with effectiveness, they did say it helped with making their issue list more complete by adding relevant points they missed.

P5: “Opnieuw, ik vond niet dat de AI hem effectiever maakte. Maar misschien wel efficiënter. Ik zit altijd met het verschil. Voor mij zijn dat twee hele verschillende termen. Het ene gaat over krijg je voor elkaar wat je wil. Nou, in allebei de gevallen. Maar bij efficiënt gaat het over hoeveel moeite doe je



ervoor. En deze haalt wat moeite weg.”

“Ik ben nu iets vollediger, want hij pakt nog een paar dingen uit... die ik zelf misschien niet meteen erbij had gehaald... maar wel relevant vind achteraf.”

P1: “Maar het is niet dat ik hiervan verwacht dat dit een heel groot impact zal hebben op mijn werk. Het is fijn dat het er is.”

### **Presentation**

P1: “Nice that point & ref are separated; don’t need headers (kopjes). I find it concise and powerful (short and sweet) like this.”

P1: “Nice dat de punten en quotes gescheiden zijn. Verder vind ik het zo eigenlijk kort en krachtig. Heb geen behoefte aan kopjes (voor de punten).”

## **C.3 Experimental set-up**

P1: “(About the client) I just want to know whether they have different or wrong conclusions based on what they saw, which I would have to rectify (during the Debrief). If that is the case, I’ll have to put in more effort to go against it. Plus, it also shows what the client is more interested in, so what I should pay more attention to (when writing the report).”

P1: “(About the client) Maar ik wil ook gewoon weten of zij misschien andere of verkeerde conclusies trekken op basis van wat ze hebben gezien. Die ik dan weer recht moet trekken. Dus waar ik dan iets harder tegenin moet gaan. En dat toont natuurlijk ook waar zij het meest geïnteresseerd in zijn. Dus waar ik extra aandacht aan kan besteden.”

P1: “I thought it (the debriefing of the simulated test day) was very easy as I did not have an actual client I had to convince.”

P1: “Nu vond ik het heel makkelijk want ik had geen klant die ik moest overtuigen eigenlijk.”

P1: “I thought it (the debriefing of the simulated test day) was very easy as I did not have an actual client I had to convince.”

P1: “Vaak zie je pas echt grote verbanden tussen inzichten of thema’s nadat je meerdere interviews hebt gedaan.”

P1: “(Regarding the use of the issue list) Usually I do use it after I’ve done multiple interviews. When certain topics are mentioned more often; then I write them down in the issue list. And I would also write down what we just discussed (during the Debrief). I would fill in the issue list during quiet moments, for example after the second or third interview.

P1: “(Regarding the use of the issue list) Die gebruik ik vaak wel hoor, als ik meerdere interviews heb gedaan. En als ik dan bepaalde onderwerpen vaker terug hoor komen. Ja. Dan schrijf ik die daar wel op. Dat wat we nu hebben besproken zou ik daar dan ook op schrijven. Ah, oké. Dan doe je dat tijdens de check-in op de e-brief? Ja, dat doe ik dan op de rustige momenten. Dat doe ik dan bijvoorbeeld na de derde interview of tweede interview bijvoorbeeld. Of als ik dan iets hoor tijdens een gesprek wat ik eerder ook al voorbij zag komen.

Dan schrijf ik dat ook even op. Of ik kopieer het even vanuit de Opslog naar de Issuelist. De Issuelist vind ik meer een beetje als een dump omgeving van dingen die ik belangrijk vind in de Opslog en die ik niet wil vergeten. Die ik wil dan graag gebruiken voor de debrief, of voor het vormen van conclusies.”

P3: “Obslogging improved my understanding of the obtained insights, a lot or not really. But the obtained insights refer to your own insights that emerge during obslogging. Isn’t that the principle of obslogging?”

P3: “Het doen van deze taak, het opsloggen, heeft mijn begrip van de verkregen inzichten helemaal niet verbeterd of ontzettend verbeterd. Maar dat zijn toch je eigen inzichten die tijdens het opsloggen naar voren komen. Dus dat is toch het principe van obsloggen? Dat is toch het principe van vergaren?”

## C.4 Obslog vs Debrief

P5: “I think this (the AI Debrief assistance) is of a whole different (higher) level than what was given in that Obslog column (referring to the AI Obslog assistance).”

P5: “Ik vind het wel interessant. Ik vind dit wel van een heel ander niveau dan wat er in dat ene kolommetje staat (referring to AI Obslog assistance)”

P7: “I quite liked the AI labels. Like that I can quickly see the ‘Context’, ‘Label’ and AI labels. Those three together give me a comprehensive view of what has been said [...] However for the issue list, the AI points are removed from their context. It just gives you ten points from the whole interview, making it difficult to check whether the point is true or not. Then I have to rely on my memory for verification. That’s where it can go wrong.”

P7: “Nou, ik vond het dus vooral met die labels vond ik het wel nice. Dat ik dan toch wel heel snel even kan zien context, label, AI-label. Dat die drie samen best wel een goed beeld geven van wat er is gezegd.”

“Maar bij de issue list is het uit de context gehaald. En staan er eigenlijk tien punten van dit zijn problemen binnen. Of dit zijn de dingen die zijn opgevallen. En dan is dan kan ik niet meer controleren van was dat nou waar of niet. Dat moet ik dan uit mijn eigen geheugen gaan halen. Daar kan dan iets misgaan”

P6: “For example here he (the AI Debrief assistance) says ‘improved communication’. But without knowing where exactly in the interview this has been mentioned, then it is difficult to find it back to understand what it is actually about/ what it specifically refers to.”

P6: “Maar bijvoorbeeld hier zegt hij -verbeterde communicatie-. Maar als ik dat even zonder te weten waar in het gesprek het is, als ik dat zo lees, dan moet ik weer helemaal erin duiken om te begrijpen van, oh ja, maar waar ging het eigenlijk ook alweer over?”

P2: “Instead of having a new tool, which would mean that I would have to do away with all the habits that I built up in the last five years, which make me so fast (in performing the test day tasks), because I want to work with AI. But

now you have found a way to incorporate it into my workflow.”

“When I got your invite I thought ‘Oh, there we go again with the umpteenth ChatGPT tool, which I’ll just put away in some list (like a bookmark for later use). The newest AI tool that is supposed to enrich my life, but actually gives extra noise.”

P2: “In plaats van dat het een nieuw iets is waardoor ik al mijn gewoontes –die ik nu eigenlijk in vijf jaar heb opgebouwd en daarom ben ik ook zo snel geworden– allemaal moet loslaten omdat ik met AI wil gaan werken. Maar nu heb je een manier gevonden om dat in mijn workflow soort van toe te passen.” Toen ik jouw uitnodiging kreeg dacht ik, oh dan gaan we weer de zoveelste ChatGPT en die. En dan heb ik ze ergens weer in de lijst. De nieuwste AI tool die mijn leven moet verrijken, maar ook weer extra ruis op levert.