Universiteit Leiden
The Netherlands

Opleiding Informatica

Usability Optimisation for Fundamental Rights Assessment of Algorithms

Hannah Gibb

Supervisors:
Prof.dr.ir. J.M.W. Visser & Prof.dr.ing. A.J. Klievink

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl
15/01/2025

Abstract

Background: Due to the rise of AI in governmental institutions, regulations are required that ensure negative effects such as unjust discrimination are limited. IAMA (*Impact Assessment Mensenrechten en Algoritmes*, translated to Fundamental Rights and Algorithm Impact Assessment), developed by researchers and government staff, analyses the effect of algorithms on fundamental rights, and will soon become mandatory by law for the public sector. Development has started of software that allows IAMA to be completed by multiple users in a well-organised manner, as part of a bachelor thesis project by Koen Bron. This software requires further improvement for users to easily understand and use it.

Aim: This thesis investigates, firstly, whether applying Nielsen's usability heuristics to IAMA Checker is effective in finding usability problems. We then examine whether the new software version developed by implementing changes to address these problems allows users to better understand it and effectively use it.

Method: To identify usability problems, we systematically review the software for each of Nielsen's usability heuristics. This is done both individually by the author and in interactive sessions with government staff familiar with IAMA. To improve the system with respect to the identified problems, we followed the Design Science Research Methodology. We prioritise addressing ambiguity and navigation issues in design implementation. We then perform evaluations using the Thinking Aloud method, and a questionnaire combining TAM, SUS and NPS methods with problem-specific and open questions. These methods focus on both the observed and perceived usability, as well as the overall attitude to the product and willingness to use it.

Results: A significant overlap is found between the usability problems found by heuristic analysis and in interactions with stakeholders. Evaluations show a positive response to the usability of the new software version, that users can easily navigate the software to complete tasks, and that user acceptance has improved since Bron's thesis.

Conclusion: We conclude that the application of usability heuristics (supplemented by user testing) to IAMA Checker is effective in finding usability problems, and that the changes implemented have improved the user experience in performing IAMA. However, the software ought to be further improved to remove remaining ambiguity and include several further items of functionality required for actual usage.

Contents

1 Introduction

An increasing number of organisations and government institutions have begun to work with algorithms and AI in order to do their work. McKinsey, for instance, has found that AI adoption in organisations has surged to 72% in 2024 from the previous approximate 50% that had remained fairly consistent between 2018 and 2023 [SSY+24]. The results these algorithms produce can have direct consequences on the lives of citizens. It is, therefore, essential to map the risks that these algorithms present and to assess whether and how it is ethical to implement them.

The Dutch government has worked together with researchers from Utrecht University to produce IAMA (*Impact Assessment Mensenrechten en Algoritmes*, translated to Fundamental Rights and Algorithm Impact Assessment) [Rij21b]. The assessment, supported by a 95-page PDF fill-in document containing explanation and form fields, helps the relevant parties consider regulations and responsibilities surrounding the use of an algorithm. To prevent discriminatory use of data and algorithms, such as occurred in the Dutch child benefits scandal (*toeslagenaffaire* [Buk20] [oK20]) from happening again, the necessary legislation has recently been passed to make IAMA mandatory for many layers of the Dutch government [Utr22]. Using IAMA can be difficult, however, due to the comprehensive array of questions and extensive explanations throughout the document.

Bron has recently started development of a web-application called IAMA Checker, aiming to support the IAMA process better than the existing PDF format [Bro24]. This software is designed both to ease novices into performing IAMA with additional explanation, and to allow for more efficient collaboration between the various parties involved. Such a well-received and useful software tool for performing IAMA would aid the implementation of IAMA and encourage its use. As described in Bron's evaluation, the proof of concept appears successful. The main point for further research described is improving the intuitiveness of the user interface.

This thesis investigates the usability problems IAMA Checker has in greater detail using Nielsen's usability heuristics [Nie94c] and implements solutions to the most important of these. The result is then evaluated to assess whether the implementations for addressing these issues has improved usability, and thus whether the application of heuristics was effective.

1.1 Problem statement

AI adoption is rising in government, with 120 AI systems currently being actively used throughout 40 government organisations [Rek24]. For 35% of these AI systems, it is unclear if they behave as they ought to, and many have not (yet) been subjected to a risk assessment. As a completed software product, IAMA Checker would allow government staff to more efficiently perform IAMAs on algorithms, thus helping to stimulate responsible and ethical implementations of algorithms in government. In Bron's thesis [Bro24], all evaluation participants entered a score of 4 out of 5 (agree) to the statement "Using this product at work would help me complete IAMA related tasks faster", and when asked how likely they were to recommend this product to a colleague on a scale from 1 to 10, the average score was a 7.8. These are among the results that show that IAMA Checker was received positively by stakeholders on the whole.

However, Bron's evaluation participants tended to agree that the user interface was lacking intuitiveness. Participants stated it was too easy to 'get lost' in the software: users lacked overview and did not know how to return to the desired page. The software also lacked useful warning or error messages at times, and clear names and placements for buttons. Moreover, comments made in evaluation discussions showed that the additional information intended for those unfamiliar with IAMA was displayed in a way that could be a hindrance to those well versed in IAMA. Fixing these usability issues, among others, is essential for ensuring IAMA Checker can provide the valuable support for IAMA users that it has the potential to do.

1.2 Research question

In this thesis, we further develop the IAMA Checker software by focusing on the improvement of usability. As such, the following research question is posed:

> **Research question** How does the application and implementation of usability heuristics to IAMA Checker improve the users' experience in performing IAMA?
>
> *Sub-question 1.* Is the application of usability heuristics to IAMA Checker effective in finding known usability problems?
>
> *Sub-question 2.* How should IAMA Checker be improved to address these usability problems?
>
> *Sub-question 3.* Does the improved version of IAMA Checker with revised interface allow users to (better) understand the structure and status of IAMA Checker? (Mental model improvement)
>
> *Sub-question 4.* Do users find the improved version of IAMA Checker with revised interface easy/easier to use efficiently and accurately? (User acceptance improvement)

1.3 Thesis overview

This thesis is a bachelor project at the Leiden Institute of Advanced Computer Science (LIACS), supervised by Prof.dr.ir. J.M.W. Visser (LIACS) and Prof.dr.ing. A.J. Klievink (Faculty of Governance and Global Affairs). This chapter contains the introduction; Section 2 discusses the background and related work; Section 3 discusses methods; Section 4 describes the results, including the design and implementation; Section 5 discusses the results and Section 6 concludes by answering the research questions and describing future work.

2 Background & Related Work

This chapter provides additional background on IAMA and usability, and discusses related research on this topic.

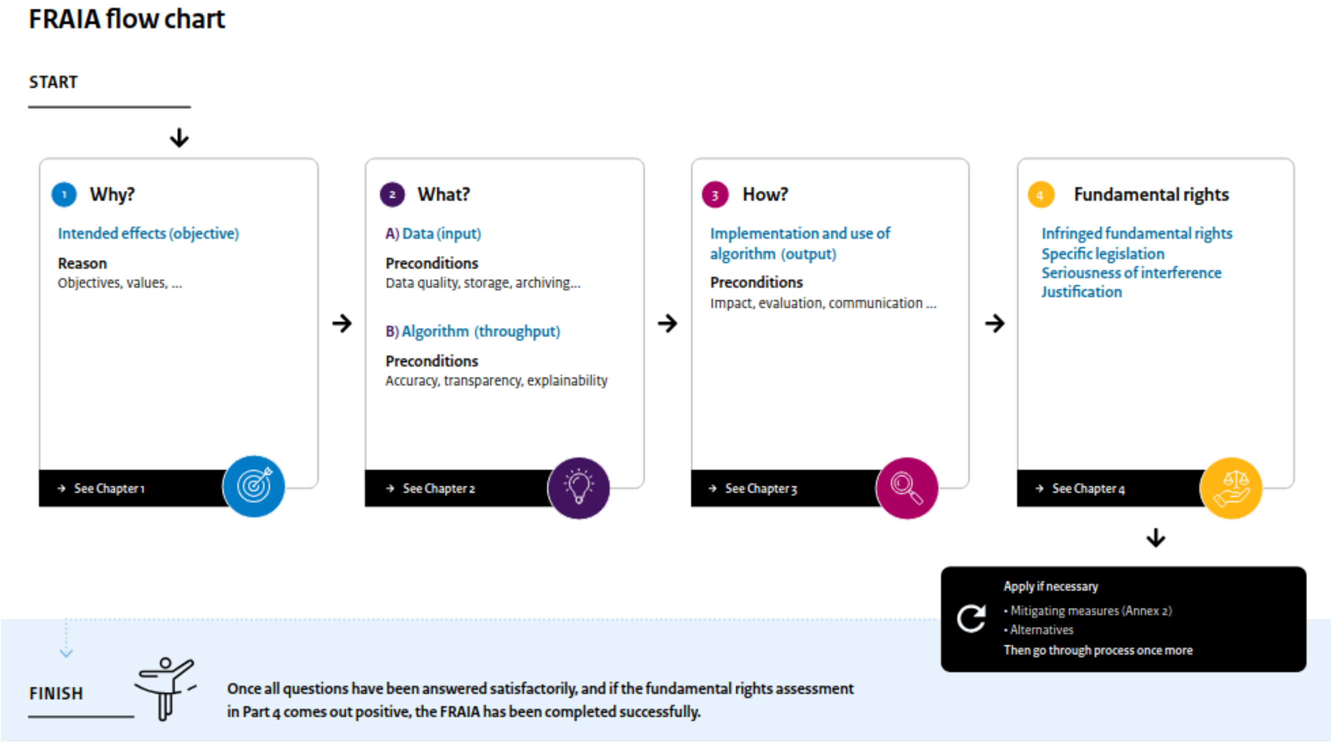

FRAIA flow chart
START
1 Why?
Intended effects (objective)
Reason
Objectives, values, ...
See Chapter 1
2 What?
A) Data (input)
Preconditions
Data quality, storage, archiving...
B) Algorithm (throughput)
Preconditions
Accuracy, transparency, explainability
See Chapter 2
3 How?
Implementation and use of algorithm (output)
Preconditions
Impact, evaluation, communication ...
See Chapter 3
4 Fundamental rights
Infringed fundamental rights
Specific legislation
Seriousness of interference
Justification
See Chapter 4
Apply if necessary
• Mitigating measures (Annex 2)
• Alternatives
Then go through process once more
FINISH
Once all questions have been answered satisfactorily, and if the fundamental rights assessment in Part 4 comes out positive, the FRAIA has been completed successfully.

Figure 1: Overview of IAMA (FRAIA in English translation) [Rij21a]

2.1 IAMA

IAMA aims to support informed discussion between relevant parties as to whether (and if so, how) to develop and implement an algorithm, in order to prevent unknown consequences, "ineffectiveness or infringements on human rights" [Rij21a]. It is recommended that various roles (if applicable) are physically present for an IAMA session (such as the project leader, legal advisor, domain expert and commissioning client) as discussions between disciplines are valuable for reflection on the use of an algorithm. A typical IAMA session involves a discussion leader, a note-taker and the other participants whose input is required. IAMA consists of four stages, as can be seen in Figure 1.

Stage 1 focuses on the intended effects of the algorithm. Questions on the reason (problem statement), objectives, public values, legal basis, stakeholders and responsibilities are asked. It ensures that the overarching details on the algorithm are clear before more technical or legal questions are asked.

Stage 2 consists of two sub-parts, focusing on the input data and the throughput of the algorithm itself, respectively. The questions asked aim to ensure that sufficiently reliable, good quality and secure data will be used, that possible biases are accounted for, that other algorithm alternatives have been considered, and that there are clear agreements on the algorithm ownership, accuracy and transparency.

Stage 3 focuses on how the algorithm output is handled: whether humans are involved in decision-making and what the effects and risks of using the algorithm could be. The algorithm output is also compared to the public values and objectives mentioned in stage 1, to ensure they align. Finally, questions are asked about the procedures, context, communication and safeguarding surrounding the use of the algorithm.

Finally, stage 4 aims to facilitate structured discussion as to fundamental rights that the algorithm could have an effect on, following seven steps. The legal context and degree of infringement ought to be determined, so that these can be weighed against the necessity and effectiveness of the algorithm. Based on this information, informed decisions can be made regarding the implementation of the algorithm, and whether any measures need to be taken against the potential risks it poses.

IAMA was developed by Prof. mr. Janneke Gerards, Dr. Mirko Tobias Schäfer, Arthur Vankan and Iris Muis for the Ministry of the Interior and Kingdom Relations (*Ministerie van Binnenlandse Zaken en Koninkrijksrelaties*). This ministry is responsible for the further development and implementation of IAMA, and has recently produced a report on the experiences of various government organisations while performing assessments on 15 different algorithms [Rij24]. This report shows that participants were often pleasantly surprised by IAMA and viewed it as a helpful tool. There were, however, problems that occured, such as difficulty involving all necessary roles or understanding when IAMA was necessary for an algorithm. The government's ICT Guild (*Rijks ICT Gilde*) and researchers from Utrecht University aim to address these issues in further development.

2.2 Usability

The intuitiveness of a user interface forms part of software's usability, defined by ISO-9241-210:2019 as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [iso19]. Usability evaluation is, therefore, an important element of developing software. Two popular methods of usability evaluation are user testing and heuristic analysis [sTLB09]. User testing involves allowing users to interact with the software and noting their impressions, comments and struggles. For heuristic analysis, evaluators (most often a group) analyse the interface using usability heuristics to find potential usability problems [Nie94a].

In the 1990s, Nielsen published ten usability heuristics for the design of human-computer interaction: (1) visibility of system status, (2) match between the system and the real world, (3) user control and freedom, (4) consistency and standards, (5) error prevention, (6) recognition rather than recall, (7) flexibility and efficiency of use, (8) aesthetic and minimalist design, (9) help users recognise, diagnose, and recover from errors and (10) help and documentation [Nie94c]. These heuristics have since been used as the basis for numerous evaluations and guidelines for specific system types, to the extent that out of 70 reviewed studies, "most of the heuristics for specific domain[s] proposed heuristics that showed obvious links with Nielsen's heuristics" [HL16]. For example, Nielsen's heuristics have been combined with web design perspectives to create a usability inspection technique designed specially for web applications [CMMT07], or used in addition to collaboration-specific heuristics for a composite evaluation [KPK$^+$10].

In this case, IAMA Checker lacks intuitiveness, clarity and easy navigation, leading to users becoming confused or lost in the software, and making mistakes. Nielsen's heuristics are directly related to these issues (the first heuristic, for instance, being visibility of system status, which is essential for users to learn the system and determine how to proceed), and are excellent at explaining usability problems [Nie94a]. They have also been constructed in order to address the most major issues that hinder a system for being usable, allowing us to target areas that will produce the best results.

3 Methods

This chapter describes our methodology, outlining our application of usability heuristics, design approach and evaluations.

3.1 Usability heuristics

While several sets of usability heuristics exist, Nielsen's is the most widely accepted, as can be seen in the frequency of its use in literature [JLR16]. The most common criticism is that these principles' generality risks missing domain-specific problems. However, in Hermawati's review of domain-specific heuristics [HL16], the question is posed "on the real contribution of heuristics for specific domains", due to the frequency of Nielsen's heuristics being used as the basis for research and lack of proper validation showing advantages and weaknesses of specific heuristics. For this reason, we chose to use Nielsen's heuristics for this thesis, supported by stakeholder feedback where necessary.

In order to effectively find usability problems with Nielsen's 10 usability heuristics, we divided the IAMA Checker software into segments that we systematically traversed for every usability heuristic. Thus, for every heuristic, after reviewing the literature on that heuristic, we (individually) moved through the software and made note of any possible problems or improvements noticed. This generated a list sorted by usability heuristic, which we later sorted by software segment.

An essential part of usability is ensuring the software "speaks the user's language" and that the user can easily understand the structure, layout and status of the software [Nie94a]. Therefore, while most heuristic analysis could be done by us alone, we also held a meeting with stakeholders in order to discuss elements of the software that are unclear from the user's perspective, or displayed too little or too much explanation. The main stakeholders are staff members of the Ministry of the Interior and Kingdom Relations who manage the development of IAMA and, in some cases, participate in IAMA sessions as discussion leaders or note-takers. In addition to supplying additional required information, we also discussed usability problems that staff members observed, constituting a limited form of user testing. This allowed for a comparison of what arose from our individual heuristic analysis and what stakeholders observe, especially since some of the stakeholders were present for Bron's work and evaluation of the prototype. Using this data, the first sub-question (*Is the application of usability heuristics to IAMA Checker effective in finding known usability problems?*) can be answered.

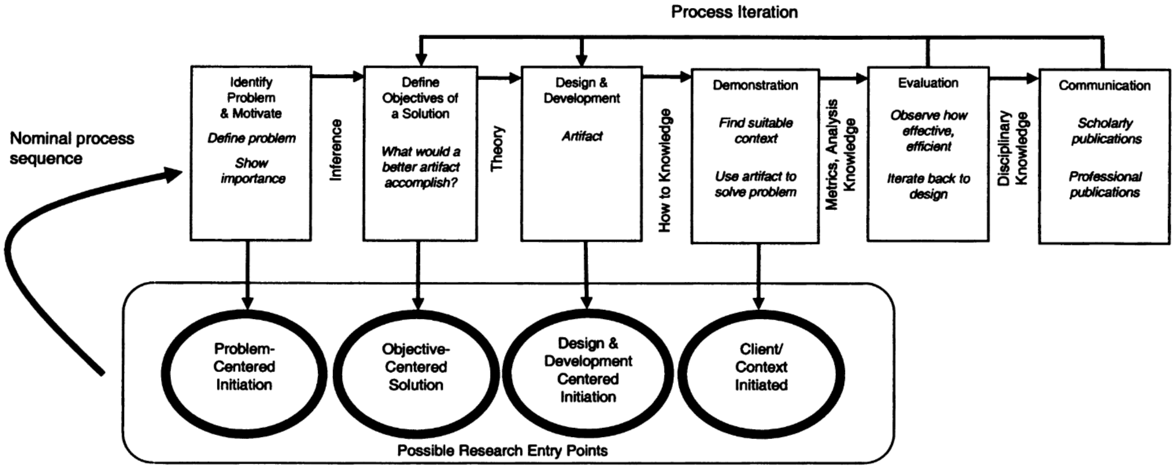

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 2: Overview of Design Science Research Methodology [PTRC07]

3.2 Design

The methodology of the user interface design and implementation uses Design Science Research Methodology as its basis, designed by Peffers et al. as a standardisation for design science [PTRC07]. The methodology entails six steps, as can be seen in Figure 2.

Bron's previous work on IAMA Checker also followed this methodology [Bro24], allowing this research to continue his work with further iterations of the design science research process. An overview of how we applied Design Science Research Methodology to the problem addressed in this thesis can be found in Figure 3. As can be seen in the diagram, the process then starts with step 2: defining the objectives of a solution. This is done by reviewing the evaluations of the existing prototype, applying usability heuristics to the software and discussing usability issues with stakeholders. This is explained in more detail in Section 3.1. Having determined the solution, the improvements are implemented in step 3: design and development. This entails the improvements being implemented and integrated into the code. This step also includes an intermediate moment of feedback and discussion with the stakeholders, to evaluate whether changes need to be made to ensure the design in process meets its objectives.

Step 4 of the design process is demonstration, in which the implemented solution is shown to solve one or more parts of the problem and address the determined objectives. This is the opportunity to show the new interface to the stakeholders of IAMA Checker. This can then be evaluated with the stakeholders in step 5, in which users navigate and use the software for IAMA so that ease of use can be observed. Following conversations and surveys can ascertain whether users understand the software structure and perceive the new prototype as more efficient. Evaluation methods are further explained in Section 3.3. These results will then be incorporated into the final thesis, which forms step 6: communication.

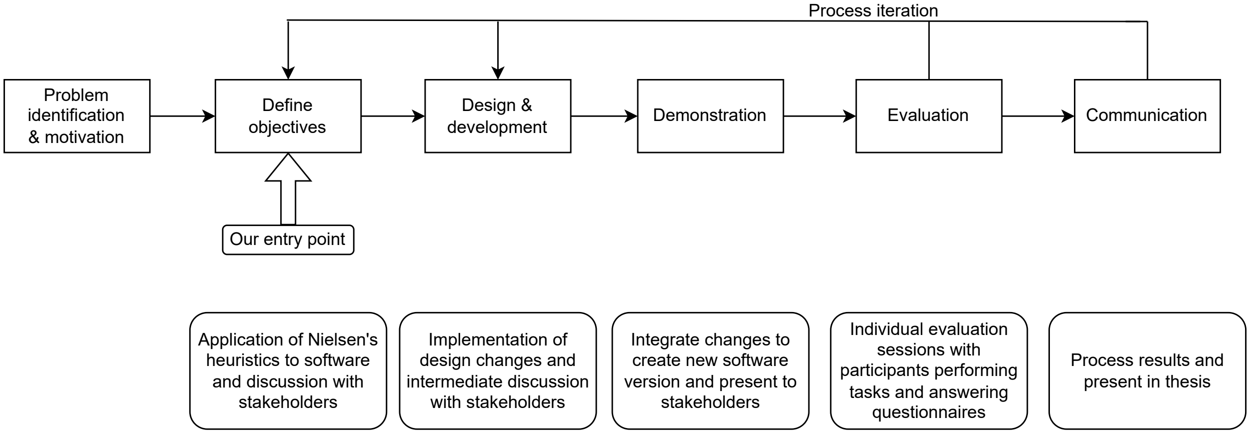

Process iteration
Problem identification & motivation
Define objectives
Design & development
Demonstration
Evaluation
Communication
Our entry point
Application of Nielsen's heuristics to software and discussion with stakeholders
Implementation of design changes and intermediate discussion with stakeholders
Integrate changes to create new software version and present to stakeholders
Individual evaluation sessions with participants performing tasks and answering questionnaires
Process results and present in thesis

Figure 3: Our application of Design Science Research Methodology

3.3 Evaluation

The evaluation of the new version of IAMA Checker aims to investigate several things: the usability and user experience of IAMA Checker, the user perception of the prototype, and whether they would want to use it in real life situations. The evaluation of our work on IAMA Checker is done with individual sessions with various stakeholders from the government's ICT Guild (*Rijks ICT Gilde*) working at the Ministry of the Interior and Kingdom Relations, researchers from the Data School of Utrecht University and an employee at the Dutch Tax Administration (*Belastingdienst*). Evaluations take place on location so that the software can be used on our own device (eliminating potential technical issues with installation) and their experiences can be observed and discussed.

Participants are first given a few minutes to explore the software on their own and comment on their first impressions, which we take notes of. Participants are then provided with a list of tasks which were constructed with the goal of allowing users to cover all functionalities and elements of the software. These tasks include logging in and out, creating a new assessment, adding an editor, answering questions and navigating between assessments. The full list of tasks can be found in Appendix B. The final task in this list is to mention any comments or other tasks they could think of, in order to facilitate discussion about what the software would require to be used in practice.

During the entire evaluation session, participants are asked to continuously think aloud and express any thoughts or comments on their experiences. As they do so, we take notes of their thoughts and sometimes respond with questions for elaboration or clarification. This usability assessment method is, according to Nielsen (author of the ten usability heuristics used in this research) "... [possibly] the single most valuable usability engineering method." Using this method allows for valuable insights with a small number of participants, as studies show that five participants can find 77-85% of usability problems of a software and running more test subjects has progressively diminishing returns [Nie94b].

After having completed all assigned tasks, participants are asked to fill in a questionnaire on their experience with the software. This questionnaire is provided on our own device and created using

Google Forms[1], which is described in more detail below. As users think aloud filling in the form, we sometimes ask follow-up questions and discuss their answers as they fill them in.

3.3.1 Questionnaire

The questionnaire consists of four sections, the first of which being six questions taken from the Technology Acceptance Model (TAM) [Dav87]. These assess the perceived usefulness of the product, which is caused by the perceived ease of use and the system design features. As these questions were also used in Bron's evaluation of the previous version of the software prototype [Bro24], these results can also be used for comparison. The questions are answerable using the Likert scale [Lik17] ranging from 1 (strongly disagree) to 5 (strongly agree) and are adapted to the context of IAMA, as follows:

Perceived Usefulness (Technology Acceptance Model):

1. Using this product at work would help me complete IAMA related tasks faster.
2. Using this product would improve my IAMA related job performance.
3. Using this product would improve my productivity when working with IAMA.
4. Using this product would increase my effectiveness at working with IAMA.
5. Using this product would make it easier to do my job when working with IAMA.
6. I would find this product useful when working with IAMA.

The second section focuses on perceived ease of use and uses the System Usability Scale (SUS) as proposed by J. Brooke [B⁺96]. This model has proven to be "robust and reliable" and gives a global view of the usability as perceived by the user. It was also used by Bron for the previous evaluation of IAMA Checker [Bro24], allowing for comparison of the results. These questions, again, use the Likert scale [Lik17] from 1 to 5 as answers for the following ten statements:

Perceived Ease of Use (System Usability Scale):

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

[1] Google Forms: https://www.google.com/forms/about/

8. I found the system very cumbersome to use.

9. I felt very confident in using the system.

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

The third section of the questionnaire was focussed on specific usability problems found with heuristics and stakeholder discussion, in order to assess whether implemented changes had positive effects. Therefore, the six questions focus on ease of navigation, understanding of the software's elements, functions and buttons, and the amount of information presented to the user. The statements, answerable using the Likert scale [Lik17] from 1 to 5 are as follows:

Problem-specific questions:

1. I found it easy to understand how the system worked.

2. I thought the layout of the system was too busy.

3. I found it easy to understand where I was in the system and how to proceed.

4. I thought too much information was presented to me while using the system.

5. I could find all the information I needed while using the software.

6. I could easily understand the buttons and parts of the software.

The problem-specific questions were followed by the Net-Promotor Score (NPS) question, summarising the user's general impression of the software prototype using the Likert scale 1-10 [Lik17]: *How likely are you to recommend this product to a colleague?*

Finally, the fourth section of the questionnaire consists of five open questions. One of these asks for an elaboration of question 6 of the problem-specific section and the others ask the participants about features or characteristics of the software that they appreciate or that require improvement. These were chosen in order to gauge which usability issues have been sufficiently resolved and which still require work. The questions are as follows:

Open questions:

1. If the answer to question 6 of the previous section [I could easily understand the buttons and parts of the software.] was not "5 (strongly agree)", what did you find hard to understand? Which parts were confusing to you?

2. What did you like about the software prototype?

3. What did you dislike about the prototype?

4. What would make it easier for you to use the software prototype (with regards to improving current functionality, not introducing extra functionality)?

5. What additional features or changes would you need to use this software prototype?

The results of these open questions are compared with each other and summarised to be used to inform discussions and conclusions from all results.

4 Results

4.1 Usability heuristics

Summaries of the findings of both the usability heuristic analysis and stakeholder feedback can be found in Appendix A. This appendix is divided into three sections, presenting the usability issues found by both heuristic analysis and stakeholders in Appendix A.1, usability issues found by heuristic analysis but not mentioned by stakeholders in Appendix A.2 and usability problems mentioned by stakeholders but not found in our heuristic analysis in Appendix A.3. This section presents some examples of these findings.

The overlap between heuristic analysis and stakeholder feedback totals 21 items. Several of these also overlapped with improvement points mentioned in Bron's evaluations [Bro24] and were among those deemed the highest priority usability problems in discussions with stakeholders. For instance, the navigation was found to be unintuitive: menus were in two separate places and needed to be merged, and it was often unclear on what page a user was located due to page titles being hardly visible. The interface was also too crowded: many bright colours drew information away from important elements, and the number of elements meant users often felt overwhelmed by what the interface presented to them. Additionally, the system lacked warning messages when users attempted to leave a page with unsaved work. The full list of usability issues in this category can be found in Appendix A.1.

The usability issues found by heuristic analysis but not by stakeholders totals 22 items. For example, successful edits in answers do not produce confirmation messages, meaning users must pay close attention to subtle changes to determine the result of their action. There is also some ambiguity present, such as in the *Klik om assessment te updaten* button (translated to "Click to update assessment"). Moreover, assessments can be deleted with two clicks without a method for recovery. The full list of usability issues in this category can be found in Appendix A.2. It should be noted that this category includes all items that were not mentioned in the initial interactions with stakeholders, regardless of whether they arose in evaluation sessions. It is possible that these issues would have been mentioned in initial discussions with stakeholders in the absence of higher priority issues. However, such investigation largely falls outside of the scope of this thesis.

Finally, the usability issues found by stakeholders but not by our heuristic analysis totals 14 items. This includes issues specific to the user domain (such as fixing the question numbering that does not match the IAMA document), which do not point to weakness in heuristic analysis but rather to our lacking understanding of said domain. However, other issues in this list do point to possible weaknesses of heuristic analysis and show the value of user discussion and feedback. This topic will be further discussed later chapters. Examples of these are questions not drawing enough attention from the user due to the location of the question title (making the essential question content harder to find and read), and the scope of the "answer history" (*vraag geschiedenis*) button being unclear due to its placement relative to other elements in the interface.

4.2 Design

This section describes the issues listed in Appendix A whose improvements were chosen for implementation in the software. The source code, documentation and issues found by our usability analysis can be found on our GitHub page: https://github.com/hjgibb/IAMA-checker.

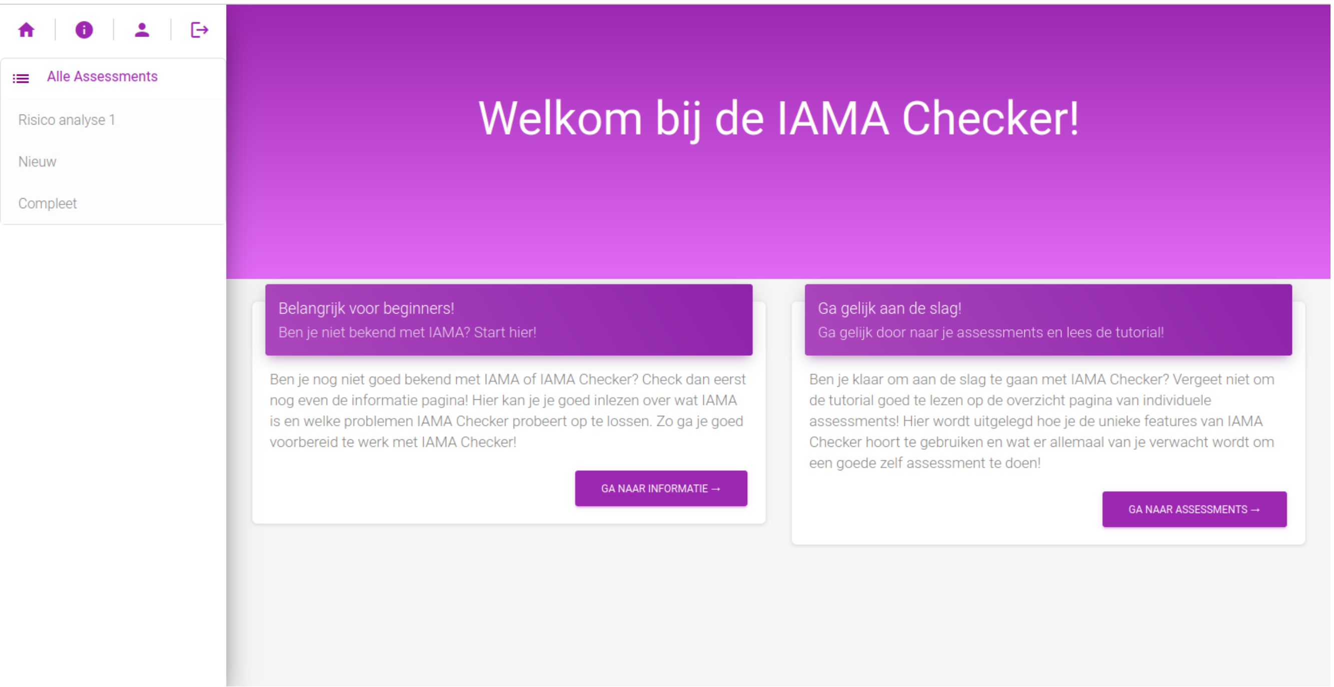
The following usability issues that were found both from heuristic analysis and stakeholder feedback (as listed in Appendix A.1) were implemented:

- Ensuring page titles are more visible
- Removing the confusing *Terug naar assessment overzicht* ("back to assessment overview") button when users are already on said page
- Improving the navigation menu and merging it to one location
- Ensuring phase 4 is properly displayed in the navigation menus
- Moving the "start assessment" button to be visible instead of hidden at the bottom of the page
- Using commonly used icons for Home and Info
- Removing ambiguous uses of the term *overzicht* ("overview")
- Removing confusing *uitputtend* ("exhaustive") functionality
- Using variation in fonts in answer history to increase readability
- Removing several cases of unnecessary and distracting bright colours

The items listed above were found by discussions with stakeholders to be among those with the highest priority. The focus is on improving the navigation menus, readability, understanding of buttons and removing ambiguity where it is present. This is supported by Bron's evaluation of the software prototype, where users found it "too easy to 'get lost'" and not know how to return to the desired page [Bro24]. Several other points in Appendix A.1 were also considered high priority, but did not get implemented due to lack of time.

The following usability issues that were found in stakeholder feedback but not in heuristic analysis (as listed in Appendix A.3) were implemented:

- Renaming statuses to be clearer and call-to-actions
- Fixing question numbering that does not match IAMA document
- Moving answer history button to be lower to ensure that the scope is clear
- Redesigning saving functionality so that marking a quesiton as *definitief* is a button, not a checkbox
- Redesigning the question page so the question itself draws more attention



Welkom bij de IAMA Checker!

Belangrijk voor beginners!
Ben je niet bekend met IAMA? Start hier!

Ben je nog niet goed bekend met IAMA of IAMA Checker? Check dan eerst nog even de informatie pagina! Hier kan je je goed inlezen over wat IAMA is en welke problemen IAMA Checker probeert op te lossen. Zo ga je goed voorbereid te werk met IAMA Checker!

Ga gelijk aan de slag!
Ga gelijk door naar je assessments en lees de tutorial!

Ben je klaar om aan de slag te gaan met IAMA Checker? Vergeet niet om de tutorial goed te lezen op de overzicht pagina van individuele assessments! Hier wordt uitgelegd hoe je de unieke features van IAMA Checker hoort te gebruiken en wat er allemaal van je verwacht wordt om een goede zelf assessment te doen!

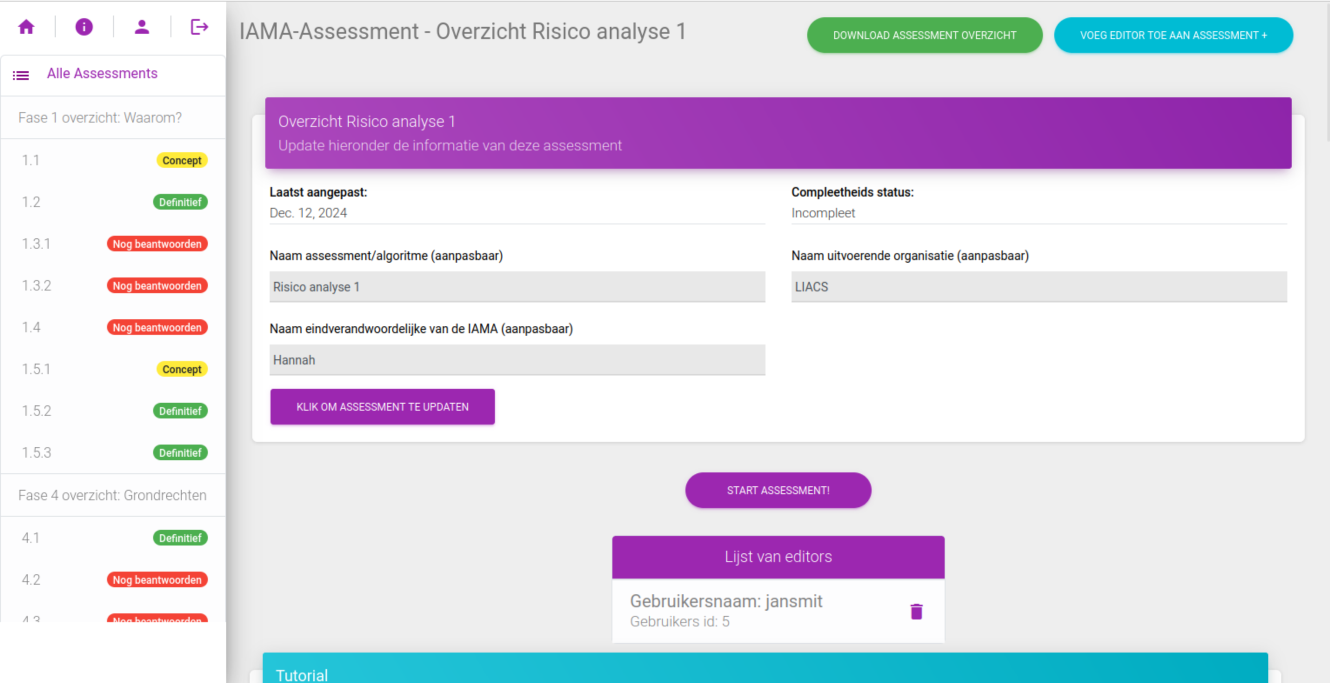
Figure 4: The Home screen of IAMA Checker, including the updated navigation menu

- Making all additional information for a question collapsible
- Restructuring phase 4 and rephrasing required professions per question to be suggestions

Other items in the list of issues found by stakeholders were considered lower priority or out of scope. No usability issues found by heuristic analysis but not by stakeholders (as listed in Appendix A.2) were considered as high priority for this research and thus not implemented.

Some of the results of these implementations can be seen in Figures 4, 5, 6 and 7. Figure 4 shows the Home page and the implemented changes to the navigation menu can be seen: while it used to contain a list of buttons for “IAMA Home”, “Assessments”, “Information”, “Account” and “Logout”, most of these buttons have been replaced by compact icons to make room for shortcuts to assessments the user has access to. This allows for fast navigation for experienced users.

Figure 5 shows the overview of an example assessment, including the updated navigation menu, now adjusted to show an overview of the specific assessment in question. While the overview of questions and their statuses used to be a separate navigation menu on the right side of the screen, this has now been merged with the left navigation bar. This ensures that all navigation is in one place, making it easier for the user not only to find where they are in the software, but also know where to find buttons needed to proceed. It also reduces the number of elements on the screen, preventing the user from perceiving it as cluttered or too busy. The confusing button to return to the assessment overview (which leads to the page already opened) has been removed and the statuses have been renamed to be clearer for the target audience, and in the case of not yet answered questions (denoted in red), act as a call-to-action. Additionally, the “Start Assessment” button has been moved to be above the editor list and tutorial, whereas it used to be at the bottom of


IAMA-Assessment - Overzicht Risico analyse 1

Overzicht Risico analyse 1
Update hieronder de informatie van deze assessment

Laatst aangepast: Dec. 12, 2024

Compleetheids status: Incompleet

Naam assessment/algoritme (aanpasbaar): Risico analyse 1

Naam uitvoerende organisatie (aanpasbaar): LIACS

Naam eindverandwoordelijke van de IAMA (aanpasbaar): Hannah

Lijst van editors

Gebruikersnaam: jansmit
Gebruikers id: 5

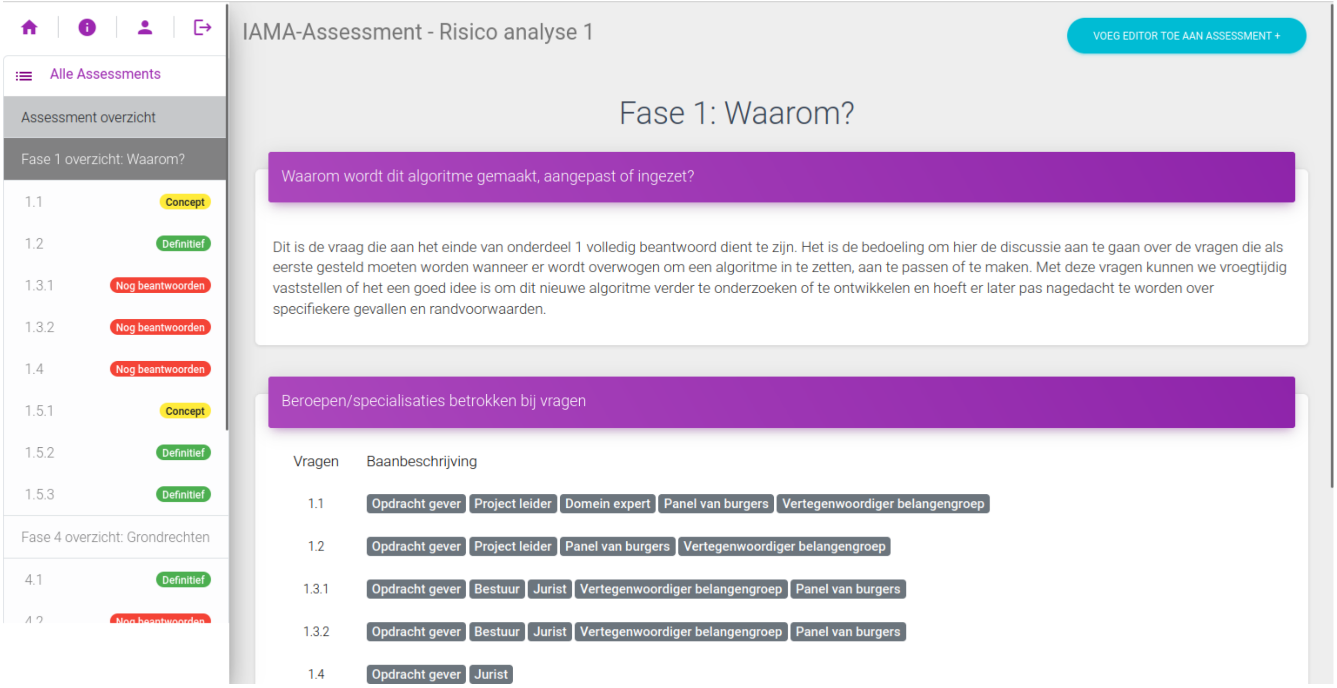
Figure 5: An example assessment overview in IAMA Checker, with updated navigation menu and moved "Start Assessment" button

the page. Finally, the tutorial header has been made blue instead of red, as to not draw too much attention away from the other important elements visible.

Figure 6 shows the overview of phase 1 of IAMA, with the updated table for which roles should give input. This table header used to be pink, but has been changed to purple in order to not draw too much attention away from the information above it. For this same reason, the badges showing the roles involved have been changed from red (required) and yellow (recommended) to grey. Stakeholders approved of the removal of this distinction, as experience with IAMA has shown that it was sometimes misinterpreted.

Figure 7 shows a question page from phase 4. This page has been restructured so that the question title is next to the question number (instead of in the purple header), the question text is in the purple header (instead of below it) and the question instructions are hidden until the grey button labelled *Vraag instructies, context en uitleg* (translated to "Question instructions, context and explanation") is clicked on by the user. This has been done to ensure the most important elements on the screen stand out, and that the user does not perceive the interface as overloaded with information. The phase 4 questions are now also properly displayed in the navigation menu on the left of the screen, to prevent confusion for the user.

Additionally, the question history button which used to be situated in the top right corner and include the text *Open vraag geschiedenis* (translated to "Open question history"), has been reduced to an icon and moved to below the question. This new position is intended to avoid confusion as to the action's scope, as all other actions that appear in the top right corner are applied to the entire



Alle Assessments

Assessment overzicht

Fase 1 overzicht: Waarom?

1.1	Concept
1.2	Definitief
1.3.1	Nog beantwoorden
1.3.2	Nog beantwoorden
1.4	Nog beantwoorden
1.5.1	Concept
1.5.2	Definitief
1.5.3	Definitief

Fase 4 overzicht: Grondrechten

4.1	Definitief
4.2	Nog beantwoorden

IAMA-Assessment - Risico analyse 1

VOEG EDITOR TOE AAN ASSESSMENT +

Fase 1: Waarom?

Waarom wordt dit algoritme gemaakt, aangepast of ingezet?

Dit is de vraag die aan het einde van onderdeel 1 volledig beantwoord dient te zijn. Het is de bedoeling om hier de discussie aan te gaan over de vragen die als eerste gesteld moeten worden wanneer er wordt overwogen om een algoritme in te zetten, aan te passen of te maken. Met deze vragen kunnen we vroegtijdig vaststellen of het een goed idee is om dit nieuwe algoritme verder te onderzoeken of te ontwikkelen en hoeft er later pas nagedacht te worden over specifiekere gevallen en randvoorwaarden.

Beroepen/specialisaties betrokken bij vragen

Vragen	Baanbeschrijving
1.1	Opdracht gever, Project leider, Domein expert, Panel van burgers, Vertegenwoordiger belangengroep
1.2	Opdracht gever, Project leider, Panel van burgers, Vertegenwoordiger belangengroep
1.3.1	Opdracht gever, Bestuur, Jurist, Vertegenwoordiger belangengroep, Panel van burgers
1.3.2	Opdracht gever, Bestuur, Jurist, Vertegenwoordiger belangengroep, Panel van burgers
1.4	Opdracht gever, Jurist

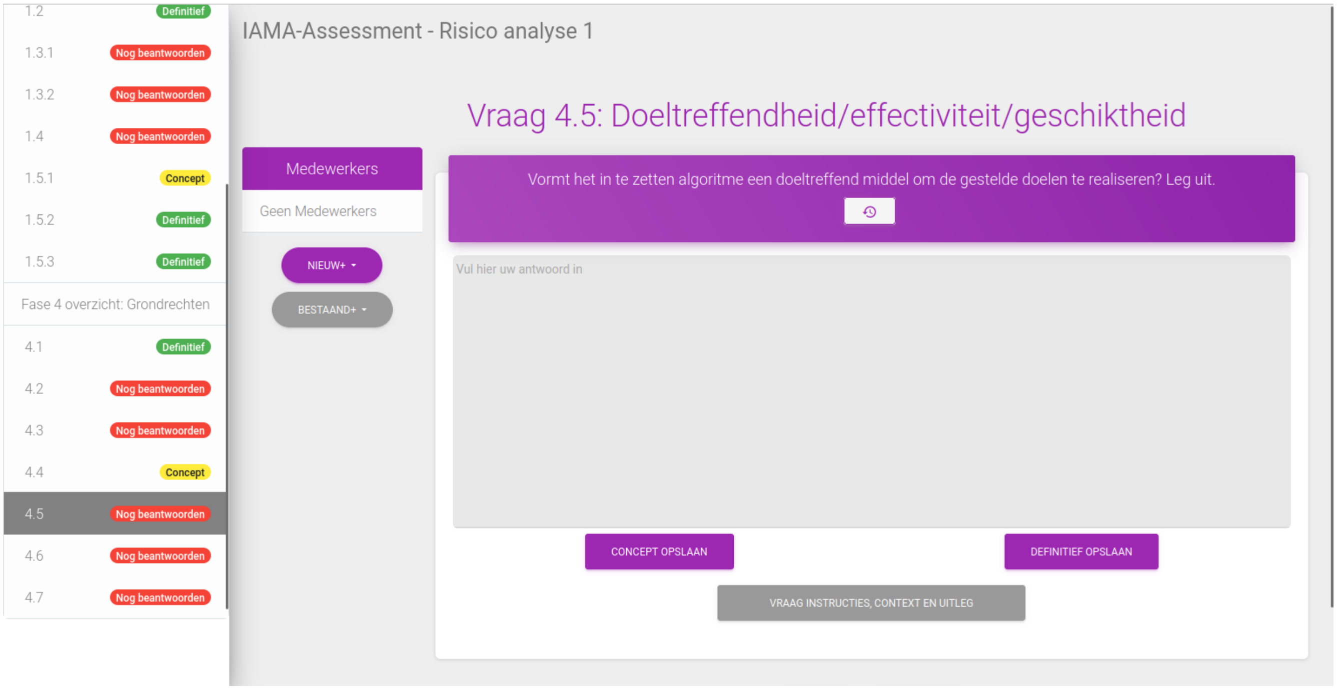
Figure 6: An example phase overview in IAMA Checker, where the badges have been changed from red and yellow to grey

assessment (instead of one question). Moreover, the saving buttons have been updated: instead of one save button and a checkbox labelled 'reviewed', a button for *Concept opslaan* (translated to "Save as draft") and another for *Definitief opslaan* (translated to "Save definitively") has been chosen in order to have more clarity and consistency with the navigation menu statuses.

Finally, as part of the design phase, we also held an intermediate evaluation with stakeholders in order to evaluate the progress made and discuss how to ensure the design meets its objectives. Here, we particularly discussed how to ensure terms used in the software are clear to the target audience , and what would be the best way of addressing some usability issues (e.g. decluttering the interface). The progress made was received positively, though stakeholders mentioned that the colour combinations currently in use could benefit from revision, that the "Dashboard" icon was unclear and that answer history would benefit from having labels for *concept* ('draft') and *definitief* ('definitive'), among others. While the "Dashboard" icon was among the points raised that have since been revised, several others unfortunately did not fit within the time constraints of this thesis.

4.3 Evaluation

This section presents all results gathered during the evaluations of the software prototype. The results for the closed questions are presented in Tables 1, 2, 3 and 4 of Section 4.3.1, summaries of responses to open questions are presented in Section 4.3.2 and a summary of observational results, including comments made during Thinking Aloud, can be found in Section 4.3.3. It should be noted that due to the lack of statistical significance of our quantitive data, open question responses, observational and Thinking Aloud data provide necessary context needed when interpreting the



Vraag 4.5: Doeltreffendheid/effectiviteit/geschiktheid

Vormt het in te zetten algoritme een doeltreffend middel om de gestelde doelen te realiseren? Leg uit.

Figure 7: An example question page in IAMA Checker, with phase 4 corrected to display properly in the navigation menu, new saving buttons and moved question history button

answers to our closed question responses.

4.3.1 Closed questions

Table 1 shows the responses received to the Perceived Usefulness (TAM) questions used. It should be noted that due to a mistake, question 4 was not included in the questionnaire and therefore lacks results. The answers to these questions are largely positive, with responses lying between "agree" (4) and "strongly agree" (5) for most questions. Question 2 is the only one to deviate from this pattern – this is discussed in Section 5.2.

These results suggest improvement from the first software version developed and evaluated by Bron, where answers were more often lying between "neutral" (3) and "agree" (4) [Bro24]. While Table 1 shows some "strongly agree" (5) responses to every question, Bron only reports receiving such responses for two of the six questions, and with a lower frequency.

Table 2 presents the results to Perceived Ease of Use (SUS) questions. While these responses are also largely positive, there are several questions in which a portion of the answers are neutral or, in the case of question 6, somewhat negative.

When comparing these results to the those of Bron [Bro24], it would appear from the responses to several questions that there has been an improvement since the previous software version. This is particularly noticeable for the following statements. For statement 1 ("I think I would like to use this system frequently"), the average answer has shifted from "neutral" (3) to lying between "agree"

#	Statement	1 (strongly disagree)	2 (disagree)	3 (neutral)	4 (agree)	5 (strongly agree)
1	Using this product at work would help me complete IAMA related tasks faster.	0%	0%	0%	(2) 33.3%	(4) 66.7%
2	Using this product would improve my IAMA related job performance.	(1) 16.7%	0%	(1) 16.7%	(3) 50%	(1) 16.7%
3	Using this product would improve my productivity when working with IAMA.	0%	0%	0%	(3) 50%	(3) 50%
4	Using this product would increase my effectiveness at working with IAMA.	-	-	-	-	-
5	Using this product would make it easier to do my job when working with IAMA.	0%	0%	0%	(4) 66.7%	(2) 33.3%
6	I would find this product useful when working with IAMA.	0%	0%	0%	(2) 33.3%	(4) 66.7%

Table 1: Results of the IAMA Checker evaluations to the TAM questions [Dav87]. Results formatted (quantity) percentage%

(4) and "strongly agree" (5). Where responses to statement 3 ("I thought the system was easy to use") previously lay between "disagree" (2) and "agree" (4), all responses are now evenly divided between "agree" (4) and "strongly agree" (5). Additionally, while Bron's responses to statement 9 ("I felt very confident in using the system") were evenly divided between "neutral" (3) and "agree" (4), all new responses are now either "agree" (4) or "strongly agree" (5).

However, for some questions, the responses appear to have become somewhat less positive, or more spread out. Namely, both statements 6 and 10 now have results that are spread slightly further from the desired "strongly disagree" (1). Discussions of these discrepancies can be found in later sections.

Table 3 presents the results to the questions specific to the navigation, information presentation and understanding usability issues found. While the responses tend more positive than negative, they do suggest that the usability problems are not yet entirely resolved.

Finally, participants were asked to summarise their experiences with IAMA Checker in a general rating question, as seen in Table 4. Comparing these results to those of Bron [Bro24] would suggest that there has been an improvement since the previous software version. Namely, the average NPS rating has risen from 7.8 to 8.7 (rounded to one decimal point).

4.3.2 Open questions

Participants were also asked five open questions about the software prototype, the answers to which are summarised below.

1. **If the answer to question 6 of Table 3 was not "5 (strongly agree)", what did you find hard to understand? Which parts were confusing to you?** Of the three participants who had answered with "agree" as opposed to "strongly agree", answers included

#	Statement	1 (strongly disagree)	2 (disagree)	3 (neutral)	4 (agree)	5 (strongly agree)
1	I think I would like to use this system frequently	0%	0%	0%	(3) 50%	(3) 50%
2	I found the system unnecessarily complex.	(2) 33.3%	(3) 50%	(1) 16.7%	0%	0%
3	I thought the system was easy to use.	0%	0%	0%	(3) 50%	(3) 50%
4	I think that I would need the support of a technical person to be able to use this system.	(5) 83.3%	(1) 16.7%	0%	0%	0%
5	I found the various functions in this system were well integrated.	0%	0%	(2) 33.3%	(2) 33.3%	(2) 33.3%
6	I thought there was too much inconsistency in the system.	(3) 50%	(2) 33.3%	0%	(1) 16.7%	0%
7	I would imagine that most people would learn to use this system very quickly.	0%	0%	0%	(4) 66.7%	(2) 33.3%
8	I found the system very cumbersome to use.	(4) 66.7%	(1) 16.7%	(1) 16.7%	0%	0%
9	I felt very confident in using the system.	0%	0%	0%	(5) 83.3%	(1) 16.7%
10	I need to learn a lot of things before I could get going with this system.	(3) 50%	(1) 16.7%	(2) 33.3%	0%	0%

Table 2: Results of the IAMA Checker evaluations to the SUS questions [B+96]. Result formatted (quantity) percentage%

#	Statement	1 (strongly disagree)	2 (disagree)	3 (neutral)	4 (agree)	5 (strongly agree)
1	I found it easy to understand how the system worked.	0%	0%	(1) 16.7%	(3) 50%	(2) 33.3%
2	I thought the layout of the system was too busy.	(3) 50%	(2) 33.3 %	(1) 16.7%	0%	0%
3	I found it easy to understand where I was in the system and how to proceed.	0%	(1) 16.7%	(1) 16.7%	(2) 33.3%	(2) 33.3%
4	I thought too much information was being presented to me while using the system.	(2) 33.3%	(2) 33.3%	(2) 33.3%	0%	0%
5	I could find all the information I needed while using the software.	0%	(1) 16.7%	(1) 16.7%	(2) 33.3%	(2) 33.3%
6	I could easily understand the buttons and parts of the software.	0%	0%	0%	(3) 50%	(3) 50%

Table 3: Results of the IAMA Checker evaluations to the problem-specific questions. Result formatted (quantity) percentage%

How likely are you to recommend this product to a colleague?									
1	2	3	4	5	6	7	8	9	10
0%	0%	0%	0%	0%	0%	(1) 16.7%	(2) 33.3%	(1) 16.7%	(2) 33.3%

Table 4: Results of the IAMA Checker NPS evaluation question. Result formatted (quantity) percentage%

that adding editors and changing who has the final responsibility (*eindverantwoordelijke*) were not where the user expected, that there was often no confirmation message of a user's action, and that parts of the software were not intuitive without having carefully read the tutorial.

2. **What did you like about the software prototype?** Clear layout; good overview of IAMAs, questions and what still needs to be done; the ability to see previous versions of answers as everything is saved in one place; and the increased efficiency of an interactive and simple program that makes it easy to delegate and include editors remotely where needed.

3. **What did you dislike about the prototype?** Language could be improved in terms of correctness, clarity and brevity. This is especially important for instructions. Titles could be made more eye-catching and possible actions from the start/home screen could be clearer. A few bugs also require fixing.

4. **What would make it easier for you to use the software prototype (with regards to improving current functionality, not introducing extra functionality)?** Participants had very varying answers to this question. The following things were mentioned: bold text on buttons, a better tutorial with more illustrative images (such as screenshots) and more mention of the status menu on the left, relocating the "next" button on questions higher up and including a "finish" or "back to overview" button on the final question. Another thing mentioned was simplifying the software by merging entities, such as editing permissions and final responsibility to be extensions of the existing "editor" entity.

5. **What additional features or changes would you need to use this software prototype in real situations?** Adding the ability to add notes, action points and to-dos separate from a question's answer, which could then be summarised on the assessment overview, was mentioned several times in answers to this question. Other things mentioned were improvements to the saving functionality (using either automatic saving or warnings if an answer has not been saved), more options for export (export to privacy-friendly PDF with no names, internal use PDF with all names listed, and an editable format such as Word) and more reassurance of security and robustness.

4.3.3 Observational data

During the thinking aloud evaluation sessions, we were able to both observe the participants' interactions with the software and gain insight into their experience with it. This subsection summarises the qualitative data collected from observations of and conversations with evaluation participants.

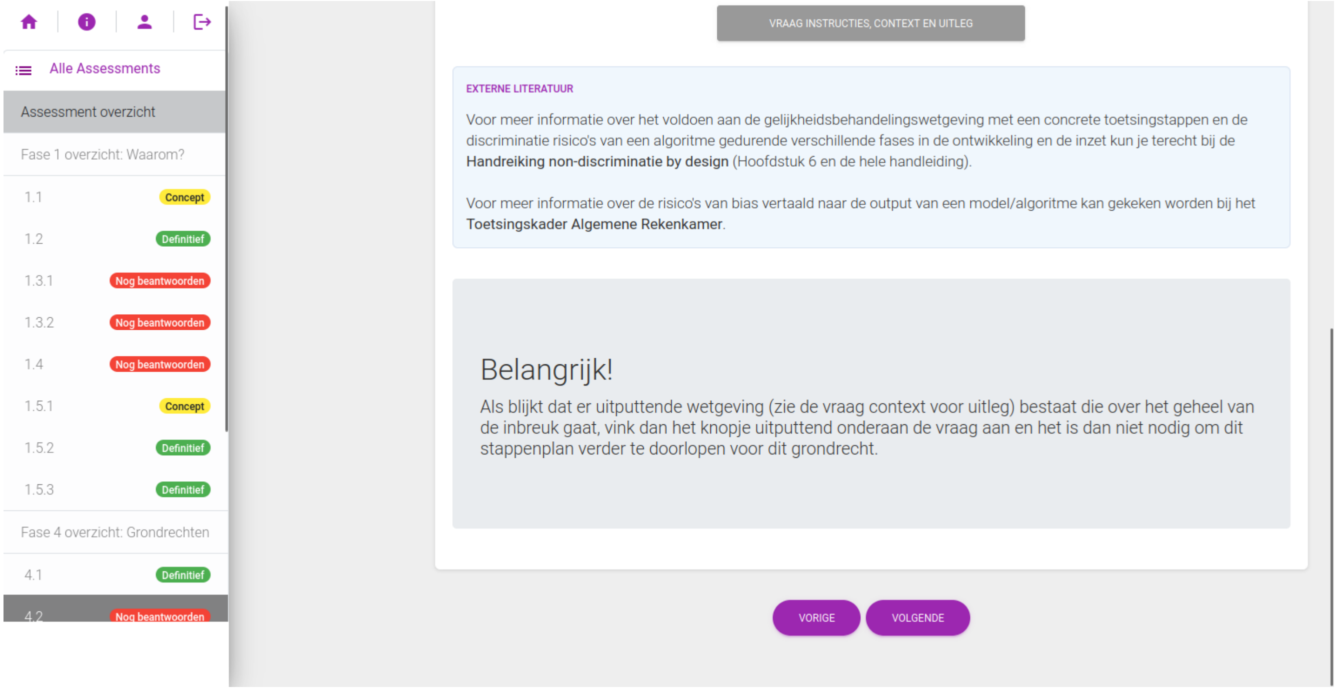
It should first of all be noted that when users had taken the time to read through the software tutorial and familiarise themselves with its functionality, very little intervention was needed to guide users through the software. Over six hour-long evaluation sessions, only twice did we help in guiding the participant to the right place in software. This is an improvement from Bron's evaluation results, where participants indicated they easily lost overview in the software and intervention was needed "sometimes" to get users back on track when stuck [Bro24]. Several parts of the evaluation instruction guide were intended to ensure users would have to navigate different parts of the software (for instance, question 11 where users have to copy an answer from one assessment to another), and these were completed successfully by all participants with no assistance from us.

However, not all observations made were positive. Certain mistakes were made by multiple participants and while they were not so much of a hindrance that they were mentioned in questionnaire answers, we recorded these as points that require improvement. Firstly, more than half of participants expected to be able to add an editor account to an assessment via the list of the editors, while the button is actually in the top right corner (see Figure 5). Secondly, there is a button visible in the navigation menu that brings the user from a question or phase to the overview of that assessment (see Figure 6). However, even when the quickest way to complete an assignment from the instruction manual was to make use of this button, only one of the six evaluation participants did. When we pointed this out to participants, they responded either by saying that the button did not appear clickable to them due to its colour, or that they did not understand from the button name what it meant.

Another mistake made by multiple evaluation participants was, when instructed to log out and in again, clicking on the "Account" icon before realising that the icon for logging out is next to it. Multiple users also believed the *Klik om assessment te updaten* (translated to "Click to update assessment") button on the overview of an assessment (see Figure 5) to be a button for retrieving edits made by other users instead of saving changes made to assessment properties: Name, Organisation and Person in charge. Lastly, it is clear from observing the participants' experience with IAMA Checker that adding editors is not currently implemented in an intuitive manner: despite instructions stating the user-ids of accounts to be added, participants repeatedly filled in usernames instead.

Besides observational data, there are several discussion points that were not mentioned in questionnaire answers. Several participants pointed out the lack of confirmation message when an answer has been saved, an editor added, or the properties of an assessment have been changed. The placement of elements on the screen was also commented on a few times: a participant expressed the thought that the blocks titled *Belangrijk!* (translated to "Important!") should be placed higher on the screen, and that the *Vorige* and *Volgende* buttons (translated to "Previous" and"Next") should be placed above the block labelled *Externe literatuur* (translated to "External literature") - see Figure 8.

Some users also expressed to have expected or required some form of text for the icons in the top left corner (Home, Information, Account, Logout) - either below the icon or as a tooltip. Finally, some additional functionality was also suggested: the ability to add attachment documents to answers so that referencing other texts is easier, the ability to archive assessments (and thus not permanently delete them), *concept* and *definitief* ("draft" and "definitive") labels on answers in question history, and the ability to update IAMAs every year or every phase of an algorithm's life cycle.



EXTERNE LITERATUUR

Voor meer informatie over het voldoen aan de gelijkheidsbehandelingswetgeving met een concrete toetsingstappen en de discriminatie risico's van een algoritme gedurende verschillende fases in de ontwikkeling en de inzet kun je terecht bij de **Handreiking non-discriminatie by design** (Hoofdstuk 6 en de hele handleiding).

Voor meer informatie over de risico's van bias vertaald naar de output van een model/algoritme kan gekeken worden bij het **Toetsingskader Algemene Rekenkamer**.

Belangrijk!

Als blijkt dat er uitputtende wetgeving (zie de vraag context voor uitleg) bestaat die over het geheel van de inbreuk gaat, vink dan het knopje uitputtend onderaan de vraag aan en het is dan niet nodig om dit stappenplan verder te doorlopen voor dit grondrecht.

Figure 8: Screen layout on an IAMA question - an evaluation participant commented that the elements at the bottom of the screen should be reordered

5 Discussion

This section discusses the results and their limitations, starting with the heuristic analysis in Section 5.1, followed by the evaluation results in 5.2.

5.1 Usability heuristics

Before discussing the results of our heuristic analysis, it should be noted that we have not exactly followed the recommended method. Heuristic evaluations are ideally conducted independently by 3-5 evaluators who are “single experts” with general usability experience but no expertise in the specific application area [Nie94c]. In our case, the evaluation was carried out by a single evaluator (the thesis author) with limited, general usability knowledge, acquired through university courses. Since our usability experience is neither application-specific nor at the “expert” level, our individual analyses may be less thorough or accurate than the recommended approach.

Consequently, when using our results to answer our first sub-question (*Is the application of usability heuristics to IAMA Checker effective in finding known usability problems?*), it is possible that the application of heuristics in the recommended manner would be (even) more effective than what our results would suggest.

Of the 43 usability problems found by our heuristic analysis (listed in Appendices A.1 and A.2), 21 were also found by stakeholders in discussions of IAMA Checker. This overlap, and the importance of these usability problems, suggests that the application of usability heuristics to IAMA Checker was effective in finding the usability problems that stakeholders deemed most urgent. This is further supported by participants in the evaluation sessions pointing out several usability issues that had previously been found by our initial heuristic analysis but not raised in initial stakeholders discussions. This overlap is further discussed in Section 5.2.

There are, however, 14 issues raised in the initial stakeholder discussion that had not been found by heuristic analysis (listed in Appendix A.3). Several of these cases can be explained by our lack of extensive knowledge of the user domain. The clearest examples of this are *"Fix question numbering that does not match IAMA document"* and *"Restructure phase 4 and rephrase required professions to be suggestions"*: the discrepancies to be fixed were quickly recognised by stakeholders due to their experience with IAMA. Moreover, most other usability problems in this list are either adjustments to ensure software elements are clearly understandable for those in the user domain, or ways to achieve heuristic 8 (minimalist design) by hiding, removing or recolouring elements of the interface. The fact that both of these categories of usability problems are present in the list not found by our heuristic analysis could, besides a weakness in the heuristic analysis, point to our lack of knowledge of the user domain and deviation from the recommended method of performing the evaluation with multiple people.

5.2 Evaluation

Moving onto the evaluation results, a significant limitation of our evaluation is the method of sampling. Of the six people who participated in evaluations, two had already seen parts of the software during stakeholder discussions. While neither of them actually used the software themselves during these discussions, their limited familiarity with the layout of the software leads to a more varied representation of novice and non-novice users. This may, however, have caused bias in our results, especially on questions pertaining to the ease of learning the software.

Additionally, five of the six participants were relatively young (under the age of 40). As also pointed out by one of the participants, this is unlikely to be representative of the user population, which is presumably more varied between the ages of 20 and 65. As it may be easier for younger people to quickly understand new software, this may also affect the reliability of the results and conclusions of this thesis.

Furthermore, it should be noted that due to the small number of six participants, the quantitative data and the comparison thereof to Bron's evaluation cannot be considered statistically significant. Therefore, in using our results to draw conclusions and answer the research questions, the observational data (from the Thinking Aloud method) and answers to open questions should weigh heavier and/or be considered in combination with the quantitative data of the closed questions.

Nevertheless, the results from Table 1 (showing the results to the TAM questions) are, excluding question 2, decidedly positive as to the perceived usefulness of the software prototype and also an improvement from Bron's results [Bro24]. The deviation of question 2 from this pattern can be explained by several factors. First of all, as Bron describes in his thesis, job performance refers to the quality of IAMA answers. This would be improved by the explanatory functionality of IAMA Checker, which those experienced in IAMA (such as the participants of both Bron and our evaluations) are unlikely to use. Secondly, these explanations can be overwhelming to the user when made always visible, due to the amount of text. We have, therefore, implemented changes to hide explanatory information under buttons, meaning the interface is clearer and less busy, but explanations are more difficult to find. Stakeholders that participated both in the initial discussion of the interface, and the evaluations, found the implementation of this decision to be an improvement. It can, therefore, be expected that IAMA Checker is successful in improving efficiency of performing IAMA, but not necessarily improving job performance and answer quality.

Table 2 (containing the results to the SUS questions) shows more varied, but still positive results. Participants express that they would like to use the system frequently (statement 1), that it was easy to use (statement 3), that they do not need the assistance of a technical person (statement 4), that they felt people could learn the system quickly (statement 7) and themselves felt confident using it (statement 9). All of these responses are decidedly positive and also show improvement from Bron's evaluations [Bro24].

However, other responses were more neutral-leaning or, in the case of statements 6 ("I thought there was too much inconsistency in the system.") and 10 ("I need to learn a lot of things before I could get going with the system."), somewhat worse than Bron's evaluations [Bro24]. This worsening can point to, besides the lack of statistical significance, a different method of evaluation (as Bron did not report the use of specific instructions or assignments in his evaluations). It could also be explained by a difference in attitude to the prototype since the system has been further developed. Participants could be treating the prototype more as a product nearing release than a proof-of-concept, leading to a more critical approach. Participants were also instructed to read the tutorial themselves with no explanation from us, whereas Bron reported having to explain the tutorial himself due to time restrictions. These factors could have contributed to the negative discrepancies in these results. Nevertheless, responses to all statements were more positive than negative, suggesting progress in the system's ease of use.

The results in Table 3 are similarly positive-leaning but also indicate a need for improvement. Most participants found the system easy to understand (statement 1) and could easily understand the buttons and parts of the software (statement 6). This indicates that the implementation of clarifying changes in the software has been successful. Results for statements 2 ("I thought the layout of the system was too busy."), 4 ("I thought too much information was being presented to me while using the system.") and 5 ("I could find all the information I needed while using the system.") suggests that the ideal amount of information being displayed by the system is not yet in correct balance. It is vital that users can find information they need (which not all participants felt they could), but presenting too much information leads to users being distracted and overwhelmed (as stakeholders expressed occurred during the initial discussion of the interface). Additionally, while data from observation and open question results indicate that user navigation of the system has improved, the responses to statement 3 ("I found it easy to understand where I was in the system and how to proceed.") indicates there is still room for improvement.

In reviewing the answers to the open questions, we notice that several improvement points mentioned were already present in heuristic analysis that had not been mentioned by stakeholders (listed in Appendix A.2): "Successful edits and saves of elements do not produce confirmation messages" is mentioned in the answer to question 1 and "Last question has no 'continue to overview' button, only a lack of a 'next' button" is mentioned in the answer to question 4. We also noticed that "'Klik om assessment te updaten' is ambiguous and potentially misleading", found by our heuristic analysis, was not mentioned in stakeholder discussion but was confirmed by several participants using the button incorrectly during evaluation sessions (as mentioned in the observational data in Section 4.3.3). These issues were likely not found by stakeholders initially due to more prominent issues ensuring they went unnoticed. The fact these issues are now more visible suggests that the issues found initially are less noticeable or sufficiently addressed. This further supports evidence that heuristic analysis was an effective method of finding usability problems in IAMA Checker.

It should also be noted with regards to the open question responses that while we have implemented changes during this thesis to make the titles of the web pages more noticeable, participants still mentioned in the responses to question 3 that they could be made more eye-catching. While many possible improvements are mentioned by users in these responses, this is noticeably one of the few issues raised that had already been (in part) addressed by our implementations. This would suggest that this issue has not been sufficiently addressed by us. It was also already known that the software lacking warning messages if a user attempts to leave a page with unsaved work is an important issue. This was, unfortunately, not addressed with an implementation due to time constraints, but is still an important issue for future work that is mentioned in the responses to question 5. On the other hand, the aforementioned issues are the only ones mentioned in open question responses that were already known or even partly addressed. This suggests that the implementations for the addressed usability problems were successful.

Finally, with regards to the observational data in Section 4.3.3, there is evidence to suggest that participants required a lot less intervention and help than in Bron's evaluations [Bro24], suggesting significant improvement in the IAMA Checker software. However, this evidence is somewhat limited by the fact that the two evaluation processes (that of Bron and us) were not identical. For instance, Bron reports that his evaluations were often rushed in the hour reserved with participants, meaning the tutorial was often explained rather than read. Bron also does not report having used a prepared list of assignments or instructions, the contents of which (if used) likely differs from ours and could have an effect on our results. These factors indicate that while comparison to Bron's evaluation looks promising, it lacks a degree of scientific control.

Nevertheless, our observational data indicates that while participants were able to navigate the software with ease, there were several discrepancies between the interface and the understanding and assumptions of users. A notable example is that users did not correctly perceive the button visible in the side menu to bring users from a question or phase to the overview on that assessment. While several users complimented the layout and usefulness of this side menu, it is clear that it could be revisited to ensure all elements are clear to users, in order to further improve navigation. This is also the case for several other buttons such as *Klik om assessment te updaten*, adding an editor to an assessment and the icons in the top left corner. These issues emphasise the need for software elements that "speak the user's language", an essential part of usability [Nie94a], which has not yet been fully achieved. This is further discussed in Section 6.

6 Conclusions and Future Work

To conclude, in this thesis we have improved the IAMA Checker software prototype by addressing usability problems. These problems were found using both input from stakeholders in the Ministry of the Interior and Kingdom Relations and the application of Nielsen's usability heuristics. Comparing the findings of these two methods enables us to reason about the effectiveness of usability heuristics on IAMA Checker. Following the Design Science Research Methodology, we then addressed those problems with the highest priority by implementing improvements to the software. This methodology also included an intermediate evaluation session with stakeholders, and a number of individual evaluation sessions where participants' use of the software was observed using the Thinking Aloud method, discussion with participants and a questionnaire with both open and closed questions.

Comparing the results of the stakeholder discussion and Nielsen's heuristics shows that heuristic analysis has been effective in finding usability problems found by stakeholders. There is a significant overlap found by these methods, and many more issues found by heuristic analysis were later found by evaluation session participants. That being said, our application of heuristics did not find all the usability problems found by stakeholders. We, therefore, conclude that while heuristic analysis done by an individual (or small number of evaluators) provides valuable information, user testing and evaluations remain vital for comprehensive usability analysis.

The Thinking Aloud and observational results indicate a considerable improvement in the user's understanding of the software structure and status: users were now able to complete most assigned tasks smoothly and without needing any help from the author, and most indicated that they found the software easy to navigate and understand. However, there is still some room for improvement. Evaluation participants remarked on the lack of feedback from the software on its status (e.g. when progress in IAMA is saved), and still found some components confusing or misleading.

Moreover, user responses received during evaluation sessions demonstrate that the improved interface allows for efficient and accurate use: participants indicated that they found the system easy to learn and use and that using it would improve their efficiency using IAMA. All users expressed intent to use the software, and that they are (highly) likely to recommend the product to a colleague. On the other hand, there is also room for improvement on this aspect. Misclicks did occur, which reduce efficiency and accuracy of use. The amount of information shown should also be reviewed, as evaluation responses were somewhat divided on whether information shown was (in)sufficient or excessive. Participant responses to open questions indicate that some software elements were not where they expected and suggested further improvements to the software layout (e.g. making titles more eye-catching and language more concise), which would improve efficiency in use.

Therefore, we conclude that our work in this thesis has indeed improved the users' experience in performing IAMA. Future work could address important usability problems that were not chosen for implementation in this thesis, such as restructuring the “employee” (*medewerker*) and “editor” entities, implementing warnings when users have forgotten to save their work, confirmation messages when work has been saved and further minimising the interface (i.e. reducing text and reviewing colour choices).

Other areas of improvement include workflow functionality (comments on text, separate fields for notes and to-dos, inviting users to review specific answers, etc.) and security/robustness (ensuring that sensitive data is secure and that IAMA progress is not lost). Extending account functionality would also be required for the product to be put to actual use (e.g. forgotten password functionality, invite links, etc.), as well as implementing phases 2 and 3 of IAMA (instead of only phases 1 and 4). Further research would also benefit from increasing evaluation sample size to permit statistical significance, including more distribution in participant age and experience, and extending to other evaluation set-ups such as a field study (i.e. using the software in a real-life IAMA session).

References

[B+96] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[Bro24] Koen Bron. Collaborative web-based tool support for iama assessments on public sector algorithms. Thesis bachelor informatica, LIACS, Leiden University, 2024.

[Buk20] G.N.J.A. Bukkems. De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag. Rapport, Authoriteit Persoonsgegevens, 2020.

[CMMT07] Tayana Conte, Jobson Massolar, Emilia Mendes, and Guilherme Horta Travassos. Web usability inspection technique based on design perspectives. In *Anais do XXI Simpósio Brasileiro de Engenharia de Software*, pages 394–410. SBC, 2007.

[Dav87] F. Davis. *user Acceptance of Information Systems: The Technology Acceptance Model (TAM).* University of Michigan, 1987.

[HL16] Setia Hermawati and Glyn Lawson. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied Ergonomics*, 56:34–51, 2016.

[iso19] Iso 9241-210:2019 ergonomics of human-system interaction part 210: Human-centred design for interactive systems, 2019.

[JLR16] Cristhy Jimenez, Pablo Lozada, and Pablo Rosas. Usability heuristics: A systematic review. In *2016 IEEE 11th Colombian Computing Conference (CCC)*, pages 1–8, 2016.

[KPK+10] Nikos Karousos, Spyros Papaloukas, Nektarios Kostaras, Michalis Xenos, Manolis Tzagarakis, and Nikos Karacapilidis. Usability evaluation of web-based collaboration support systems: the case of cope_it! In *Knowledge Management, Information Systems, E-Learning, and Sustainability Research: Third World Summit on the Knowledge Society, WSKS 2010, Corfu, Greece, September 22-24, 2010. Proceedings, Part I 3*, pages 248–258. Springer, 2010.

[Lik17] Rensis Likert. The method of constructing an attitude scale. In *Scaling*, pages 233–242. Routledge, 2017.

[Nie94a] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, page 152–158, New York, NY, USA, 1994. Association for Computing Machinery.

[Nie94b] Jakob Nielsen. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, 41(3):385–397, 1994.

[Nie94c] Jakob Nielsen. *Usability Inspection Methods.* John Wiley & Sons, New York, NY, 1994.

[oK20] Parlementaire ondervragingscommissie Kinderopvangtoeslag. Ongekend onrecht. December 2020.

[PTRC07] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007.

[Rek24] Algemene Rekenkamer. Focus op ai bij de rijksoverheid. https://www.rekenkamer.nl/publicaties/rapporten/2024/10/16/focus-op-ai-bij-de-rijksoverheid, October 2024.

[Rij21a] Rijksoverheid. Fundamental rights and algorithms impact assessment. https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms, July 2021.

[Rij21b] Rijksoverheid. Impact assessment mensenrechten en algoritmes. https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes, July 2021.

[Rij24] Rijksoverheid. Iama in actie - lessons learned van 15 iama-trajecten bij nederlandse overheidsorganisaties. https://www.rijksoverheid.nl/documenten/rapporten/2024/06/20/iama-in-actie-lessons-learned-van-15-iama-trajecten-bij-nederlandse-overheidsorganisaties, June 2024.

[SSY$^+$24] Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. The state of ai in early 2024: Gen ai adoption spikes and starts to generate value. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai, May 2024.

[sTLB09] Wei siong Tan, Dahai Liu, and Ram Bishu. Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics*, 39(4):621–627, 2009.

[Utr22] Universiteit Utrecht. Tweede kamer stemt in met verplicht gebruik impact assessment mensenrechten en algoritmes. https://www.uu.nl/nieuws/tweede-kamer-stemt-in-met-verplicht-gebruik-impact-assessment-mensenrechten-en-algoritmes, April 2022.

A Appendix: Usability problems by method

A.1 Overlap Nielsen's heuristics & stakeholder feedback

Total 21

Heuristic 1: visibility of system status

- Page titles are hardly visible/noticeable
- 'Completion status' could be more effectively displayed as a percentage or progress bar
- 'Terug naar assessment overzicht' button is visible even if user is already on said page
- Navigation menu requires improvement: menus should be merged to be in one singular location and made to be collapsible to avoid too much information
- Phase 4 is not displayed properly in the navigation menus

Heuristic 2: match between the system and the real world

- Difference between "medewerker" and "editor" unclear
- Difference between "nieuw" and "bestaand" medewerker unclear

Heuristic 3: user control & freedom

- "Start assessment" button & editor list should not be hidden right at the bottom of the assessment page, underneath the tutorial (also falls under heuristic 7: flexibility and efficiency of use)

Heuristic 4: consistency and standards

- Commonly used icons can be used for Home, Info, Back, Save, Edit, Download (also falls under heuristic 6: recognition rather than recall)
- Term 'overzicht' used in too many locations, causing ambiguity
- Distinction and different layout of owner and editor assessment tables unclear

Heuristic 5: error prevention

- No warning message if user leaves page without saving answer
- 'Uitputtend' functionality very unclear

Heuristic 6: recognition rather than recall

- Adding an editor should be easier: using e-mail rather than user ID, and with a list of current editors & other user accounts visible

Heuristic 8: aesthetic & minimalist design

- Colour combinations are overwhelming and should be reviewed

- Answer history difficult to read due to use of same font in whole window
- Info page is overwhelming: titles do not draw attention, too much information without bullet points or bold fonts to accentuate what is most important
- Answers in history contain no formatting, paragraphs or text hiding in case of long answers
- Interface is generally very busy, colourful and cluttered

Heuristic 9: help users recognize, diagnose and recover from errors

- Trying to definitively save an empty answer should give a clear error

Heuristic 10: help & documentation

- Tutorial lacked clear topic buttons (where irrelevant information is hidden) or a search function

A.2 Nielsen's heuristics (but not mentioned in stakeholder discussion)

Total 22

Heuristic 1: visibility of system status

- Left navigation menu lacks use of highlighting to show where in software the user is located
- Creating assessment redirects to said assessment, which is unclear
- Successful edits and saves of elements do not produce confirmation messages
- Highlighting of text box when selected could be clearer
- 'Close' buttons on pop-ups do not react very much to hovering over them
- Last question has no "continue to overview" button, only a lack of a "next" button
- Tabs in law clusters pop-up do not give clear indication of current open tab

Heuristic 2: match between the system and the real world

- "Klik om assessment te updaten" is ambiguous and potentially misleading

Heuristic 3: user control & freedom

- Info page has no buttons except navigation menu, can feel a bit like a dead-end page
- Assessments can be deleted with two clicks and no way to recover it if done mistakenly
- No option to revert an answer back to a previous version
- No undo/redo (or other common texting editing) buttons

Heuristic 4: consistency and standards

- Some terms used interchangeably: ‘close’/‘sluit’, ‘onderdeel’/‘fase’, ‘updaten’/‘opslaan’
- Colour consistency: purple used for colour of titles and clickable buttons, similar dual usages for other colours can be found
- “Terug naar wetten overzicht” button responds differently from all other menu buttons

Heuristic 7: flexibility and efficiency of use

- Software lacks bookmarking functionality so shortcuts can be added to home screen (to replace tutorial, for instance)
- Sorting/personalising assessment list not possible
- Ctr-S shortcut for saving could be implemented

Heuristic 9: help users recognize, diagnose and recover from errors

- Failed answer save or broken connection should produce a clear error message
- Certain URLs lack intuitive redirects to the most likely intended page

Heuristic 10: help & documentation

- Tutorial help could be spread over multiple pages, for instance “in context”
- Tutorial could be rewritten to have clearer steps

A.3 Stakeholder feedback (but not found in heuristic analysis)

Total 14 (of which 11 align with heuristics)

- “Editors” and “medewerkers” are ideally one entity that includes both listing employee names and optionally linking employee accounts (aligns with heuristic 2 and 7)
- Rename statuses to be clearer, and call-to-actions (aligns with heuristic 2)
- Fix question numbering that does not match IAMA document (aligns with heuristic 2)
- Answer history button should be lower to ensure it is clear that the scope is different than the button next to it (aligns with heuristic 4)
- Marking a question as “reviewed”/“definitief” should be a button, not a checkbox (aligns with heuristic 4)
- “Medewerkers” should be added per assessment and then adjusted per question, instead of per question (aligns with heuristic 7)
- Fields for making a new assessment and managing “medewerkers” can be collapsed unless clicked (aligns with heuristic 8)

- Too many bright colours that draw attention (where not always needed) (aligns with heuristic 8)
- Questions do not draw enough attention due to to the location of the question title (aligns with heuristic 8)
- Remove required professions from question page (aligns with heuristic 8)
- Make all additional information for a question be collapsible under a button (aligns with heuristic 8)
- Information displayed ought to be rephrased by an IAMA-expert for a good balance between accuracy and succinctness
- Having the answer to question 4.1 visible on all subsequent phase 4 questions would be useful, same for the answer to question 1.2 on 4.4
- Restructure phase 4 and rephrase required professions to be suggestions (based on stakeholders' recent IAMA experience)

B Appendix: Instructions Final Evaluation of IAMA Checker

This document includes several tasks that can be executed in IAMA Checker in order to get a complete picture of the user friendliness. In doing so, it would be helpful to think aloud as much as possible. If something is unclear, feel free to ask questions.

1. To start, if you are not (very) familiar with IAMA and/or IAMA Checker, please read the introductory information by clicking on the info icon in the top left corner, or press "go to information" (*ga naar informatie*) from the Home page.
2. Log out and in again with the username *participant* and password *iamachecker*.
3. Create a new assessment with the name *Nieuw* and your own name and organisation.
4. Look at the overview of the assessment and read the tutorial so that you understand the functionalities of IAMA Checker.
5. When you feel ready to continue, we are going to add an editor. At the moment, there are no other accounts (editors) with access to this assessment. Add "hannah" with user-id 2 to this assessment.
6. Make sure you know what your own user-id is.
7. Question 1.1 can be answered by you. However, question 1.2 needs to be checked by a colleague and thus cannot yet be definitively saved. Fill in dummy/nonsense answers and save them in the appropriate manner.

8. We are now going to leave this assessment as it is. There are other assessments that you have access to. Find the one that is not yet complete and fill your own name in as *eindverantwoordelijke* (translated to "person in charge").

9. The questions in this assessment that are saved as draft (*concept*) need to be read and confirmed by you.

10. For question 4.4 (of the current assessment), you need the answer to question 1.2. Copy this answer and save it as the answer to 4.4. This answer needs to be checked by a colleague before it can be definitively saved.

11. The answer to question 1.5.1 needs to be the same as the answer to question 1.5.1 in the assessment named *Compleet*. Review the answer versions (in question history), choose one to copy and save it in the other assessment.

12. Add a new employee (*medewerker*) to question 1.5.1: "Jan Smit" who is a data advisor for the Ministry of the Interior and Kingdom Relations.

13. Also add Jan Smit's account as an editor (user-id 5).

14. Download the PDF summary of this assessment.

15. Are there other tasks or comments that you can think of regarding IAMA or IAMA Checker software? Name these, or perform them.