Introducing Asynchronous Remote Usability Testing in Practice: An Action Research Project

Jonna Pedersen, Malene Sørensen, Jan Stage, Rune Thaarup Høegh

Introducing Asynchronous Remote Usability Testing in Practice: an Action Research Project

Jonna Helene Holm Pedersen[1], Malene Sørensen[2],
Jan Stage[3] and Rune Thaarup Høegh[4]

[1]JCD A/S, Systemvej 6, DK-9200 Aalborg SV
[2]Capgemini Danmark, Skanderborgvej 234, DK-8260 Viby
[3]Aalborg University, Department of Computer Science, DK-0220 Aalborg East
[4]Nykredit, Fredrik Bajers Vej 1, DK-9220 Aalborg East
helene.pedersen16@gmail.com, malenesoerensen1993@me.com,
jans@cs.aau.dk, rh@nykredit.dk

Abstract. Asynchronous remote usability testing is a usability evaluation method where users and evaluators are separated both in time and space, and users are directly involved in producing the evaluation results. This paper reports from an action research project on introduction of asynchronous remote usability testing in the IT organization of a bank. The IT organization had extensive experience with traditional usability evaluation but was interested in introducing asynchronous remote usability testing into their repertoire of evaluation methods. The project was initiated with an intensive two-month collaboration followed by a six-month maturation phase. The process of introducing the method and the motivations of the IT organization for introducing the method are described in detail. The findings indicate that the conceptual understanding of usability evaluation and the engagement of the users are crucial for the outcome of the evaluation, and the tool and materials used for the evaluation was an obstacle for high-quality results.

Keywords: Methods for HCI, Usability evaluation, Remote asynchronous usability evaluation, Action research.

1 Introduction

Usability evaluation is an established discipline in modern software development. One of the most commonly proposed evaluation methods is referred to as the user-based think-aloud method, which is usually carried out in a dedicated laboratory. This traditional method has significantly influenced usability evaluation research for several years. In software development practice, there are mixed opinions about the user-based think-aloud method, with the main concern being resource demands, both in terms of calendar time and work hours for the evaluators.

There have been various efforts to meet the demands of industry practice by developing evaluation methods which require less resources. Remote usability testing (RUT) was introduced in 1994 as a less resource-demanding method, e.g. [20, 22]. RUT aims to have users and evaluators located in different places, communicating over the

internet in a way that is similar to the traditional user-based think-aloud method. The separation in space reduced the logistical challenges; the evaluators and users were not separated in time though. This has later been denoted as synchronous remote usability testing (SRUT) [1].

A related method was suggested a couple of years later. It was originally called critical incident reporting where the users are separated from the evaluator in both time and space, and report the usability problems they experience [6, 21]. The evaluators receive all the critical incident reports from the users and based on that they generate a list of usability problems. This has later been denoted as asynchronous remote usability testing (ARUT) [6].

Some research activities have inquired into the qualities of SRUT and ARUT, and some researchers have also compared the two. SRUT has many qualities in common with the traditional user-based think-aloud method, while ARUT identifies significantly fewer usability problems in comparison to the former two [1]. On the other hand, ARUT is a more cost and time effective approach in comparison to the traditional method and SRUT [5]. This is due to the reduced costs associated with traveling and logistics [3], as well as the scalability of the approach allowing evaluators to add more participants with little added cost [4]. The scalable nature means that the test can be administered to a larger variety and group of participants, as location is not an issue.

A main challenge of ARUT is the quantity of the results, as the method has been found to identify fewer usability problems in comparison to its counterparts [1, 5]. It also lacks the qualitative data that can be collected by evaluators with traditional evaluation and SRUT [14]. In the original form, ARUT is also more time consuming for the users who participate in the tests as they have to report usability problems in addition to working with the system [1, 5, 16]. However, there are more recent forms of ARUT that do not impose this task on the users.

While an increasing body of research inquires into the qualities of remote usability testing methods, the amount of research that focuses on the use of SRUT or ARUT in software development practice is severely limited; for ARUT there are very few exceptions, one being [15]. On a more general level, this illustrates a strong critique of usability research [18]. Practitioners who want to introduce ARUT in an IT organization need advice and experiences to facilitate the process, but the research literature provides very limited support in this respect.

This paper reports from an action research project which explored how ARUT was introduced into an IT organisation that relied on traditional usability evaluation methods. The action research study was carried out in collaboration with the internal IT organisation of a national bank, and focused on activities which were conducted to support the process of introducing ARUT in organisational practice. ARUT was chosen because the IT organisation wanted to explore how far they could go in reducing the evaluator efforts. The aim of this paper is to assist practitioners who intend to go through a similar process.

In the following section we present research literature that relates to our research question. Then we introduce the research method used for the action research study. Next, we present the findings of our study of the introduction of ARUT in the IT organisation, and whether it was successful and can be implemented in the organisation.

Lastly, we discuss the results in terms of introducing ARUT in an organisation, unexpected findings and our contribution to current research before concluding on our study by addressing our research question.

2 Related Work

There is a limited but increasing body of research on RUT. Some of this is defining various remote methods. A different stream of empirical work demonstrates that RUT is an effective approach [1, 24, 25]. Many of these collected only qualitative data, except [1].

ARUT has been defined and discussed in various contexts. It has been defined and discussed in relation to SRUT, RUT and traditional usability testing [5, 6, 12]. Castillo distinguished between asynchronous and synchronous remote usability testing [6] and clearly defined the different aspects of the two approaches in relation to remote user testing.

Bruun et al. discussed three ARUT methods and compared them to a traditional usability approach; they found that ARUT was more time effective but found fewer usability problems. Martin et al. compared the cost and time spent on traditional lab-based usability testing to ARUT [12]. They conducted testing using both approaches, with the lab-based testing taking place in a university setting. The results indicated that asynchronous remote testing could be considered a comparable alternative to lab-based testing, as it showed that remote testing had used less days of direct evaluator/consultant involvement, resulting in a cheaper cost per problem [12].

Research that compares SRUT and ARUT is much more limited. Andreasen et al. presented the results of a systematic empirical comparison of three remote usability testing methods and a lab-based think-aloud method [1]. The three remote methods consisted of both synchronous and asynchronous methods. The results indicated that synchronous remote testing had produced equivalent results to the lab-based testing. Results from the asynchronous testing condition were less positive with identification offewer usability issues, while spending more time completing the tasks assigned. Andreasen et al. still considers the approach to be worthwhile, as it frees the expert evaluators from a considerable amount of work and enable(s) collection of user data from a large number of participants" [1]. The efforts required by the evaluators per usability problem identified is lowest with ARUT. Thus SRUT identifies more usability problems compared to ARUT, but it also requires more effort both in total and per problem [1, 5]

Varga has also compared SRUT and ARUT but in terms of the qualitative experiences of the test subjects involved. Again, it is concluded that both of the methods have advantages and disadvantages, thus it is not possible to point to one of these methods as being superior to the other [23].

The research of ARUT has been focused on finding strengths and weaknesses of ARUT [5], however, this has all been research-based, in academic settings. There is a distinct research gap regarding the introduction and use of ARUT in industry practices. To our knowledge, there are no empirical studies of the introduction of

remote usability testing in a software organisation. According to Wixon, this illustrates a large gape between research and practice [18].

Nørgaard and Hornbæk [19] and Reeves [17] wrote articles examining the practical application of UX practices by industry professionals. Nørgaard and Hornbæk examined how UX professionals conduct and analyse the think aloud method. They found that the industry application differed greatly from the theoretical research, showing that the professionals adjusted the method to fit their specific needs [19]. An article by Reeves presented the results of an ethnomethodological study on how UX practitioners produce findings in usability testing within a design consultancy [17]. The results indicated that usability findings were produced as observers analyse the test while it was still unfolding. It challenged current views of usability findings as something that is "there to be found". The study suggests refinement of the current definition and understanding of usability evaluation to better reflect how usability issues are in practice [17]. These two articles highlight the importance of researching in industry as well as it may differ from the theoretical research.

3 Methods

In this section we present the research approach and methods applied in our action research study. We begin by presenting the research approach taken, before describing the collaboration and approach of the usability test.

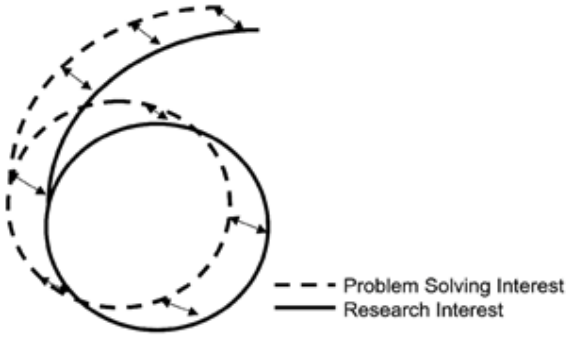

Fig. 1. Problem Solving Interest and Research Interest [13].

3.1 Research Approach

Action Research represents a duality between practice and theory. It is a research approach committed to "the production of new knowledge through the seeking of solutions or improvements to "real-life" practical problem situations" [13]. One of the distinguishing characteristics of action research is the active involvement of the researcher in the study itself [2]. "The action researcher is viewed as a key participant in the research process, working collaboratively with other concerned and/or affected

actors to bring about change in the problem context" [13]. This involves having researchers providing new knowledge, while the problem owner has the contextual knowledge required to understand the situation [13]. Action research is partially consisting of the needs and competencies of the people involved, with a key feature being the willingness to share and learn from one another [13].

In this project, we adopted Mckay & Marshall's approach to action research as consisting of two separate cycles "superimposed on each other" [13]. They state that action researchers have a dual purpose of not only having to think of their research interests but also consider the problem-solving interest within the project collaboration [13].

We divided our study accordingly; with our problem solving interest focussing on support to the introduction of ARUT in the IT organization we collaborated with, while the research interest focussed on the exploration of challenges related to the introduction of ARUT in an organisation already relying on traditional usability testing methods.

3.2 Research Collaboration

We conducted our action research study in a two-month collaboration with an internal development team in the IT organization of the bank. The development team was using SCRUM and carried out their work in two-week sprints. Their sprint deliveries would typically be additions or changes to graphical user interfaces of the applications for which they were responsible. Their interest in the project was to experience ARUT as a means to improve the quality of these sprint deliveries.

Their development process included developing drawn prototypes before investing in software development. The drawn prototypes were typically designed in collaboration with key end-users. These key end-users involvement often meant that the resulting end product was well received by the broad user group. However, the team sometimes experienced that end-users not involved in the designs, felt that significant needs were not met. This could for instance be feedback from end-users who used the applications for slightly different purposes than the involved key-users. In order to accommodate this problem, the team wanted to utilize ARUT as means to get feedback from a broader set of end-users, possibly even all end-users in their organisation.

They also saw ARUT as opportunity to gradually improve the quality of their prototypes. If significant usability problems were found in the ARUT, the team would improve the design based on the findings, and re-evaluate it in another ARUT similar to the process in the RITE-method [27].

Due to the nature of their agile development approach, the teams perspective was to focus on identifying the most significant usability problems per ARUT iteration, rather than investing in finding all possible usability problems in a more costly traditional usability test. They would also focus their efforts on addressing critical usabillity problems, over minor or cosmetical ones, since they continously had to prioritize their effort

The aim of the collaboration was therefore to introduce the organisation to the concept of ARUT while determining the value of the approach in an agile industry practice.

The problem-solving interest of our research focused on conducting an asynchronous remote usability testing of a new internal support application for a national bank. The organisation used this first ARUT as a way to gain insight into, how to further expand the use of ARUT in the organisation.

The usability test was conducted using Preely and focused on the user friendliness and usability of an application by evaluating key use cases in the application. Preely was chosen as tool, since it allows for the use and import of graphical prototypes from the teams typical design tool (Figma), and it supports easy replacements of a frame (a part of the design) as a design is gradually improved. Preely therefore supported the purpose of iteratively testing slightly changed prototypes. Other tools were also considered prior to the selection of Preely, but they were discarded for various reasons ranging from, not having the desired functionality, to failing to live up to the banks security standards regarding data being stored abroad.

Preely was also considered to be the most accessible tool for the end-users, since it did not require the end-users to setup anything on their computers. They would only be required to click an invitation link. Similarly, no video or sound recording would be stored, so the end-users could conduct the ARUT from their desks in the office without being concerned about GDPR-issues.

Communication. Due to the outbreak of COVID-19, the collaboration took place using online communication tools such as Skype and Microsoft Teams. This allowed us to conduct meetings with our contact person; conduct walkthroughs of the application, plan testing and share results.

3.3 Participants

Three types of participants participated in this study. The internal development team, participants in the ARUT, and the authors of this paper.

The development team consisted of 10 members located in three cities in two countries. It was a full stack development team with members specialising in frontend and backend development, business analysis, UX design and application architecture. Their main focus was development of applications for internal use in the bank. They developed and maintained the application the usability test was conducted on.

Throughout the collaboration, the efforts were coordinated with the team through one specific team member. The results of the collaboration were presented for, and discussed with, the entire team.

For the usability testing, we recruited end-users of the application from within the bank. The test users were 12 members from a support team of mortgage advisers. The support team's role was to help financial advisers, calling in from partner banks, with mortgage lending cases. In this study they helped us collect data related to the usability of the application.

The authors of this paper all participated in the study. Three authors were researchers from the university. They were key participants in the research process, and contributed to the research data by conducting and documenting the application of ARUT within the organisation.

The last author was employed by the bank, and a member of the development team. In the research collaboration, his focus was to align the efforts with the organisation's goal, and to provide access to the organisation. He formulated the user cases based on interviews with the end users, and he was the primary recipient of the usability test result. During the maturation phase, he disseminated the initial results of the collaboration in the organisation, and he used the results of the usability test to redesign areas with significant usability problems.

3.4 Procedure and Setting

We divided the procedures between our research and problem-solving interests. For our problem-solving interest we focused on testing the internal application remote and asynchronously. The application had been released as a pilot release to the team of mortgage advisers, and was still being adjusted by the development team according to the feedback given by two pilot users. The research interest covered the application of asynchronous testing as well as the comparison between ARUT and the traditional testing methods usually implemented within the organisation.

The usability test was conducted using an online remote testing platform used for remote users and usability testing, Preely. It allows for the creation of interactive prototypes using screenshots and/or imported prototypes from other prototype development applications. In our case, we developed a clickable prototype using screenshots of the pilot release and a list of key use cases provided to us by the development team.

The prototype was built using Preely's internal prototype builder. It was based on use cases and screenshots provided by the development team, as well as on a walkthrough of the application itself. When completed, it was sent to the development team for feedback and corrections. This resulted in an additional development session in which we made adjustments based on the feedback. After being approved, it was sent to the test users who could then begin testing.

The use cases were constructed by interviewing two end-users about typical scenarios in their daily work. The interviews were also backed up by logged data about support calls. In total, 6 tasks were formulated to represent the use cases. The tasks were set up in Preely and all tasks were presented for the individual end users in the usability test sessions. The tasks can be seen in Table 2.

Besides the usability test, we conducted interviews with test users from both the development and the support team. These were based on Kvale and Brinkmann's interview techniques [10] and addressed their experience with the approach, as well as previous experience with traditional usability testing methods.

3.5 Materials

The test users received an email containing a short guide, introducing the test and testing procedure before receiving the test itself. The guide explained how testing would be conducted, as this would be the first time that they had to conduct testing asynchronously.

3.6 Data Collection

We collected data in three different areas covering results from the usability test, data on the application of ARUT in the organisation, and data on how usability testing is currently conducted within the organisation.

Testing took place over the course of 2 weeks and resulted in 12 responses. While previous research showed examples of ARUT being conducted over the course of 24 hours [15], we would not have been able to replicate this approach. Due to the test user's inexperience with ARUT, which meant they needed more time for testing and familiarizing themselves with the concept, we found that additional time was needed. Other factors influencing the timespan were employee and bank holidays, which cut the actual time spent on testing down to 6 days.

Initially the test included a short guide introducing the test users to the concept of asynchronous usability testing. However, feedback from the users showed the guide to be lacking in its explanation and it was therefore revised to include more information on the approach and on how to identify usability issues within the software prototype. While this provided some discrepancies in the data, it minimised the user's confusion and allowed them to better identify and comment on any usability issues within the application.

Other issues included an initial lack of participation from the users as we only received 4 responses during the first week of testing. We sent out reminders and had the department manager involved to obtain as many responses as possible. This resulted in responses from all 12 department members and enough data to identify the most crucial usability issues within the application.

The end results from the usability test consist of 12 responses from all invited test users. These were collected and analysed through Preely. During the users task solving, Preely recorded the user interactions and recorded the time spent on each task. After each task Preely would prompt the users to rate their experience as well as give them the opportunity to provide additional textual feedback. After each individual tests session, Preely automatically constructed an interactive test report with information about completed tasks, given ratings, click-heat maps, and given feedback.

Aside from data collected through Preely, we conducted follow-up interviews with 2 of the users who had participated in testing.

We documented the development and testing process throughout all phases of the usability test by the diary approach presented by Jepsen et al [7]. The diary was used to monitor the time spent on developing, conducting and analysing the usability test, as well as to comment on our experience of applying ARUT in the organisation. This allowed for deeper reflection on the approach and on our research goals.

Data covering the application of ARUT in the organisation consisted of 10 diary entries describing and documenting the entirety of the development and testing process, as well as 10-minute follow up interviews with 2 test users. The interviews mostly focused on the user's testing experience and suggestions for improvements of the approach.

Lastly, we collected data on how usability testing would usually be conducted within the organisation. This consisted of an interview with a development team member in which we discussed their current testing approach, focusing on the results, time spent on testing, number of users, etc.

3.7 Data Analysis

We analysed data in two different ways; content analysis and analysis of metrics. The content analysis was implemented whenever there was textual data that needed to be analysed and sorted. The analysis of metrics was solely based on the usability test and the metrics provided through Preely.

We used content analysis for the interviews and the diary approach. The interview with the development team member was transcribed and sorted based on topics discussed during the interview. This included the tested use cases, time spent on testing, number of users, etc. These categories were later used to compare the current approach and the ARUT approach which we applied.

While we were not able to record the interview conducted with the test users, we based the data on notes taken during the interviews. These were sorted under each interview question and used to support any usability problems, as well as comment on the user experience.

The content analysis of the diary approach was done by categorising the experiences we had chronologically and thematically. We also categories the diary entries based on the development phase. This analysis was then used in the results to describe the application of ARUT and find usability problems of the testing tool.

The analysis of the usability test results was based on metrics provided in the interactive test report generated by Preely. The analysis took place immediately after testing was completed We identified usability problems by analysing the user's ability to reach the end of a task, as well as analysing their interactions while working with the task. We also examined their comments, the time it took to complete a task, and the scores the users gave after having completed a task. The usability problems were assessed based on tasks which had low scoring, a lower than an 80% completion rate, or if negative comments were given. The heat maps and interaction metrics were used to determine if a user would need to explore the prototype a lot in order to complete a task.

The usability test also elucidated certain usability problems that relate to the usability testing tool. This occurred as users often commented on their experience with the tool. These tool related usability problems were found based on the users' comments and our experience with developing the prototype.

4 RESULTS

We present the results focusing on the initial introduction and application of ARUT within the organization; How was testing conducted, what data was collected using the approach, what was our experience with ARUT and lastly how did the organization adapt and implement the approach as a part of their existing testing procedures.

4.1 ARUT Process

In this section we will present the results of the initial introduction and application of ARUT within the collaborative organisation. These are mostly based on entries made in the diary, as this allowed us to keep track of the time spent on individual tasks, as well as recall and reflect on our experiences with the testing process.

Usability Testing. On average, the test users would spend 13.25 minutes completing all the tasks. This number was however influenced by one outlier who spent 49 minutes idle before exiting the test. This inevitably affected the average time spent per user, as it would otherwise have been approximately 8 minutes per user.

Follow-up Interviews. These lasted an average of 10.5 minutes and discussed their experience testing asynchronously; problems they faced within the prototype, what they believed were the strengths and weaknesses of ARUT and how to improve the approach for future testing. The results were analysed and served as basis for future improvements of the approach, as well as to support any assumptions related to the users experience with asynchronous testing.

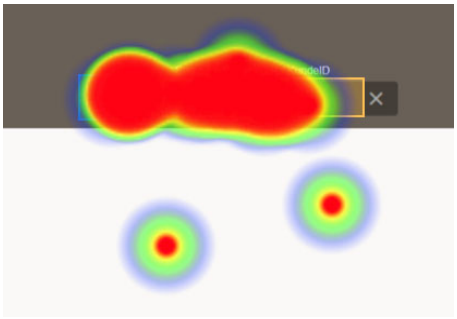

Fig. 2. Heatmap.

4.2 ARUT Results

By utilising the metrics available through Preely, we had access to plenty of information and statistics related to the users behavior and interaction with the application. This included heatmaps, action statistics, paths and clicks etc. In the following sections we

will present the results of our testing, focusing on the value of the available data-collection metrics.

Heatmaps. The heatmap provided us with a collected overview of all clicks, swipes and scrolls performed on each screen within the application. It showed areas of the UI that serve as hotspots for user interaction and was a useful tool when identifying areas that gained unintended attention from users searching to complete a specific task. This helped determine the cause for users not being able to complete a task, as it showed what part of the UI the users were focused on.

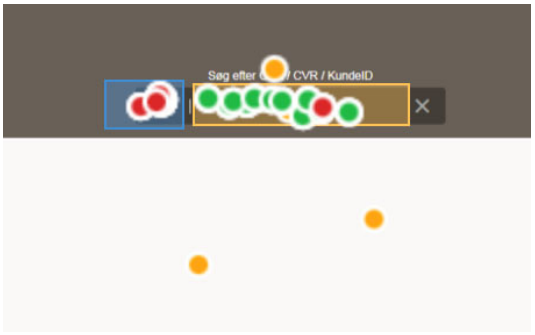

Fig. 3. Action Overview.

Action Overview. While the heatmap served as a collected overview of the user's interactions with the UI, the action overview provided more details in terms of the specific actions and order of actions the user would attempt when solving a task. It allowed us to monitor what the users were clicking on, even when those elements were not programmed to perform any actions within the prototype. This was useful for recognising any mismatch between the users expectation of how the application would work versus its actual functionality and setup.

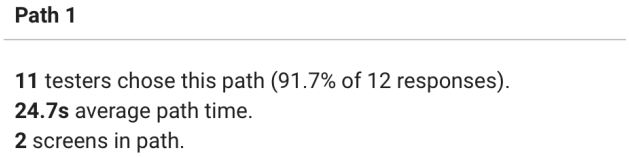

Fig. 4. Path Overview.

Paths Overview. The paths overview showed the different paths the users had taken while completing each task. This was the least useful overview, as the prototype mostly consisted of simple and linear paths that allowed users to find relevant information in a

quick and efficient manner. However, while we did not use this overview very much, we recognise the value of this in a more complex application, that would require multiple paths to reach certain information.

User Experience. The users' experience was measured through a ten-point scale and comments section included at the end of each task. The ratings and user comments represented how the users felt about the tasks and allowed them to express their thoughts on any issues in their own words.

The average rating of each task (as seen in Table 1) expressed the users experienced solving each task, with examples of comments supporting the average user scores. While most of the tasks received a positive score, some didn't work out as well in an asynchronous setting, with Task 5 & 6 receiving the lowest average scores of 0/10 and 5.5/10. In both cases Users expressed confusion over either the application or the test itself.

Table 1. Overview of task ratings.

Task	Avg. Rating	User Comments
1. You are getting a call from a consultant at the bank. The consultant is asking about customer xx (customer number xx). What do you do?	7.36/ 10	"We would usually use the customers cpr. number (in the search engine)." "It was easy and accessible"
2. The consultant tells you that the customer has requested to change the payment date. Has the change been registered in the application?	6.78/10	*"A little confusing as I could not get down to the last page before I was asked to give feedback"* *"It was easy to figure out, but I did have issues finding where I should say yes or no to the task"*
3. The consultant has asked you to confirm that the customer has received all of the relevant documents for this change.	9/10	- No User Comments -
4. The customer wants to ensure that the correct price has been used for the payment of the current loan. Which price has the customer been informed of?	6.6/10	*"Quite easy to find"* *"Was in doubt if I should click in the task text or on the payment price. The latter gave the result"*

5. You realise that there is a mistake as the user interface and document do not contain the same information. The call is over, and you decide to report the mistake.	0/10	*"I did not know that we could send an error report directly in the application"*
6. The technical support calls about the same case. The consultant has contacted them as the case has not been approved for release. What is the reason for this?	5.6/10	- No User Comments -

The overall results of the usability test indicated that 4 of the 6 predefined use cases could be completed by the users in a satisfactory manner. The two remaining use cases were considered unsuccessful and would therefore be examined in further detail. Based on the feedback, and the data recorded in Preely, the graphical user interface related to the two use cases were redesigned and once again presented to the end-users. These changes were well-received by the users and were therefore implemented in the application, in this research collaboration's maturation period.

4.3 ARUT Experience

Throughout testing, we experienced some different challenges that influenced the initial introduction and application of ARUT within the organisation. In the following section we will present these challenges and discuss how to resolve them in future implementation.

Test User Training. The users who partook in testing had no prior training or experience with ARUT before introduced to the concept through our research. While we included a written guide explaining the approach, most users seemed confused and found it difficult to identify usability issues on their own.

The organisation's current testing approach has users conduct usability testing in the presence of a moderator, who can support and guide them. This is not possible with an asynchronous testing approach and users were therefore confused in terms of what to look for and comment on. This issue could however be resolved by training the users and giving them a proper introduction beforehand - either through in-person training or an introduction video. This would require some dedicated time for training but would result in easy asynchronous testing in the future.

User Engagement. We initially saw a low engagement from test users to partake in asynchronous testing. This could be due to ARUT not being as committal as traditional usability testing, which requires users to attend testing in person. In our case, users were instructed to do testing in between other tasks, which might have influenced how committed they felt. We did see an increase in responses after the department manager encouraged the users to participate, which then resulted in the 12 responses we ended up receiving.

Another reason could be the timing of testing, as it was conducted during a busy work period, which might have resulted in the users prioritising other work over testing.

Usability Testing Tool. We did identify some issues related to the testing tool based on user feedback and our own experience working with it. These issues mostly pertained to some confusing reporting functionalities, lack of input from essential keyboard keys and issues related to the drop-off function of the test. These caused additional confusion for the users, who commented more on issues experienced with the testing tool instead of issues related to the usability of the application and prototype. These issues can however be resolved by working around the tool's weaknesses and by training the users beforehand.

4.4 Maturation Phase

The initial collaboration lasted two months, followed by a six month maturation phase. The results from the collaboration phase was considered a success by both the involved team and the organisation. The results contributed to a further investment in a broader introduction of ARUT in the organisation.

In the maturation phase, the organisation has obtained additional licenses for the Preely platform, and invested in training of UX designers and interaction designers in using Preely. Additional agile teams were being introduced to ARUT and a usability test strategy for the use of ARUT was formulated and implemented in the organisation.

Based on the experience from using ARUT on internal applications and with internal users, ARUT was also extended to be used in customer facing applications. The banks' customers are now being invited to participate in ARUT through a user panel.

The front-end team that participated in the study have employed a student worker to help them quickly set up prototypes in Preely, and are now in the process of incorporating asynchronous remote usability evaluations in their agile development process.

5 Discussion

In this section we present the unexpected findings which we encountered during the research. We discuss the findings from the organisations perspective and their actions in the maturation phase. We also discuss our findings in the context of current remote usability testing research.

5.1 Other Findings

Experience vs. Prototype. The test users commented on the experience of the test rather than the prototype of the system. They often provided comments on how they would have worded or designed tasks. They also commented on the usability of the Preely platform, or the material they had been given to guide them in the test. The end-

users suggestions include using vocabulary that the users are accustomed to, providing an extensive guide and having a greater understanding of the users and their work tasks.

While all of this feedback is relevant to mature the ARUT process in the organisation, it was also a source of noise in the concrete usability test. In order to improve the "signal to noise"-ratio, we have compiled a list of suggestion for the implementation of ARUT in organisational practice

Suggestions for introduction:

(1) The users will require a thorough introduction to the approach and tools applied, in order to avoid any confusion. This could come in the shape of either in-person training or an introductional video.
(2) The hypothesis being tested should be short and precise rather than extensive, This allows for more specific feedback on single features.
(3) The diction/language used within the task descriptions should match that of the users, not only to ensure an emulation of realistic use cases, but also to minimise confusion.
(4) The prototype should be pilot-tested with a group that mimics the demographic of the final users. This allows for any rough edges to be smoothed out before the prototype is deployed for full-scale testing.
(5) The users should have access to the development team in order to ask questions throughout testing, to make sure the testing effort is not halted due to technical or practical problems.

These suggestions come out of our action research study which was focussed on ARUT, which is why we only relate them to that approach. However, some of them, may be more generally applicable to remote usability testing in practice.

5.2 Introduction of ARUT in the organisation

In the following, we will discuss the findings related to the introduction of ARUT from the organisations perspective. Again, some of these may be applicable to other approaches than ARUT, but we only have empirical data on ARUT.

ARUT as a part of an Agile Process. The agile development team that participated in the research collaboration, found that the ARUT approach fit their needs. Their goal was to assess whether ARUT would allow for them to gain user feedback about small additions or changes to a graphical user interface in a fast way and with a low investment. The materials used for the test were created with less than two days work. The test was open for 14 days in Preely, but in reality only 6 days were used by the test users. It is further expected the amount of days needed for a test can be reduced even further with additional planning and coordination. Additionally ARUT contributed to finding two use cases that the users could not complete in the application in a satisfactory manner.

With these results in mind, the agile development team has concluded that the ARUT approach qualifies as a testing approach, that will allow for them to get user feedback on a sprint to sprint basis in an agile setting.

Training of Test Users. During the follow up interviews with User 1 and User 2, they both stated that they would have preferred more guidance throughout the usability test. They also stated that training the test users beforehand would have been a good idea as well. User 1 stated “introducing us to the testing program before would help”. If the department would like to implement this process again, training and offering more support for the usability test would improve the overall experience and results.

Seen from the organisation's perspective, the initial problems of introducing an ARUT approach to internal employees, was manageable. Although the platform itself proved to be an obstacle to the initial users, the obstacle was made smaller by adjusting the instructions. Several of the end-users later said that they saw the benefit of using an ARUT approach, as this approach would allow them to both give feedback on the application they work in on a daily basis, while also being able to fit a 10-15 minut test session in between other tasks. They also expressed that while they did have some initial problems with the platform itself, they expected them to be smaller in the future.

Training of the test users is a one time investment for internal users. The bank is also expanding the use of ARUT to external users, such as the bank’s customers using various apps, homepages or home banking solutions. In this case, the organisation might not be able to provide similar training to the individual customer. For that reason, the findings related to the instructions, training material and access to support during the test period, is especially relevant, as they otherwise risk customers having a bad experience with the test. This might impact the customers opinion on the bank itself.

5.3 Relating Our Results

In this subsection we discuss the contribution of our findings in the context of ARUT research.

This research examines the implementation of ARUT in an IT department to examine its efficiency and efficacy, filling the lack of research regarding the practical application in industry practices. Previous research conducted by Martin et al [12] and Andreasen et al [1] discusses the implementation of ARUT in comparison to traditional lab-based usability testing. The findings indicated that ARUT identifies fewer usability problems. While our results supported these findings, we also identified other factors influencing the implementation and results of ARUT in industry practices; The amount of found usability problems is less important than finding the most severe problems in a fast way, the importance of proper user training, having a good understanding of the context of testing and having a high user engagement level.

Nørgaard and Hornbæk found the differences in the theoretical discussion and practical application of the think aloud method. Our research found that several of the strengths and weaknesses of ARUT in theory did not align in practice as UX professionals have different goals when it comes to UX tests. For example, we did find fewer usability problems, but that was also because the test was targeted towards one

hypothesis; the purpose of the test was not finding many problems but finding the most significant problems within the 6 use cases. Reeves' study focused on the identification of usability findings in industry practices [17] and described how usability practitioners would work together to locate usability findings during testing. While we were not able to observe users as we conducted ARUT, we were able to analyse individual results as soon as they were submitted. This allowed us to compare results early on and adjust the test as needed.

5.4 Action Research Experiences

The research method applied in this paper is action research. The philosophical foundation of action research rise from hermeneutics, existentialism, and phenomenology [26]. There are four criteria for evaluating action research [26].

First, the researchers must intervene into the subject under study. The success of the research depends on engagement rather than detachment, Data are collected with participant observation, and this develops the empathy, the values exchange, and the role reversals that make researchers' knowledge useful and accepted by the subjects. In the research described above, the researchers intervened into the development activities of the IT organization.

Second, the project must be collaborative, and the subjects must be dynamically involved in determining the directions of the project. The researcher's role is not one of prediction in a passive world, but one of making things happen in an interactive world. In the research, the IT organization had a strong say in the design of the research. In particular, the IT organization decided that we should use ARUT and not SRUT, and they decided that we should work with an approach where the focus was on finding a few key problems rather than all usability problems.

Third, the knowledge goals should be interpretive and framed as "understanding" rather than "explanation." Although the theoretical constructs under empirical testing are complex and multivariate, they gain scientific usefulness as the conceptual "point of departure· for intervention in other settings; i.e. the theory must be interpreted and adapted in order to achieve construct validity in each new organizational setting.

Fourth, the action research project must yield a solution to the immediate problem situation. Action research develops learning from experience, which should be disseminated within the organization and published to the scientific community. This learning can lead to further action and major positive effects in diverse organizational settings. In our research, we managed to develop a solution for the IT organization.

6 Conclusion

In this article we have presented the results of a two-month collaboration with a national bank in which we introduced and conducted an asynchronous remote usability testing with the aim of determining the value of the approach for usability testing in the IT organisation. We conducted usability testing using a remote testing tool previously described by usability practitioners and implemented user-based work tasks to ensure a realistic testing experience for the test users [14] [8].

We conducted ARUT with 12 users from the internal financial support team. The test was based around 6 use cases supplied by the IT organisation. It had a structured guide for the users and all data was collected and analysed through an ARUT tool, Preely. While we were able to find relevant usability problems related to 2 of the 6 use cases tested, we found that there were several aspects that would have to be changed to maximise the efficiency of the test. This included taking the test users and their experience more into consideration during the development and planning stages. This was evident from the confusion experienced throughout testing, as test results improved when providing more guidance on how to complete individual tasks. However, we conclude that ARUT is a viable addition to the IT organisations repertoire of evaluation methods.This is supported by the fact that the IT organisation has chosen to move forward with further ARUT.

Limitations of this study included the lack of a pilot test before the launch of the actual prototype. Usually prototypes are tested with similar demographics before being tested with the final group. We tested the prototype and the link with a member of the development team, but we did not pre-test with the end-users to ensure the effectiveness of our final usability test. This could have caught any issues with the test before it was sent to the final test user. This research was conducted during the international Covid-19 health crisis, and therefore governmental restrictions also affected this research process. We were not able to meet with the development team or the usability test user in person. This is not required for ARUT testing; however, the follow-up interviews could have been conducted in person and possibly allowed for different insights.

For future research, it would be interesting to conduct an ARUT based usability test within this organisation again, following the steps outlined in the findings. The results from this research would allow for a more effective testing session and could therefore expand on the findings of this research. The implementation of ARUT could also be tested within several organisations, to compare the two workspaces and how easily ARUT can be implemented and how easily employees adapt to a new testing approach. This would allow for broader data on the general implementation of ARUT rather than the specific implementation of ARUT within one organisation.

References

1. Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. *Conference on Human Factors in Computing Systems - Proceedings*, 1405–1414. doi: 10.1145/1240624.1240838
2. Avison, D., Lau, F., Myers, M., & Nielsen, P. (1999). Action Research. *Commun. ACM*, 42, 94–97. https://doi.org/10.1145/291469.291479
3. Bartek, V., & Cheatham, D. (2003). Experience remote usability testing, Part 1: Examine the benefits and downside of remote usability testing. *Developer works* 2, 1–8.
4. Bastien, J. M. (2009, apr). Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, 79(4). doi: 10.1016/j.ijmedinf.2008.12.004
5. Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. *Conference on human*

factors in computing systems – proceedings (pp. 1619–1628). New York and NY: ACM Press. doi: 10.1145/1518701.1518948

6. Castillo, J. C., Hartson, H. R., & Hix, D. (1997). *Remote Usability Evaluation at a Glance* (Tech. Rep.). doi: 10.1145/286498.286736
7. Jepsen, L., Mathiassen, L., & Nielsen, P. (1989). Back to thinking mode: Diaries for the management of information systems development projects. *Behaviour & Information Technology - Behaviour & IT*, 8, 207–217. https://doi.org/10.1080/01449298908914552
8. Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (1996). "Quick and Dirty" usability tests. In *Usability Evaluation in Industry* (pp. 107–114). Retrieved from https://books.google.dk/books?id=ujFRDwAAQBAJ
9. Kjeldskov, Jesper; Skov, Mikael B.; Stage, Jan (2018): Instant Data Analysis: Conduction Usability Evaluations in a Day. Department of Computer Science, Aalborg University.
10. Kvale, S., & Brinkmann, S. (2008). *Part II: Seven Stages of an Interview Investigation. In Interviews: Learning the Craft of Qualitative Research Interviewing* (2nd Ed., pp. 97–290). Thousand Oaks, California: SAGE Publications, Inc.
11. Lewis, J. R. (2004). Usability Testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics methods* (pp. 1267–1312). John Wiley & Sons. doi: 10.1201/9780203489925
12. Martin, R., Shamari, M., Seliaman, M., & Mayhew, P. J. (2014). Remote Asynchronous Testing: A Cost-Effective Alternative for Website Usability Evaluation. *International Journal of Computer and Information Technology*, 03.
13. Mckay, Judy & Marshall, Peter. (2001). The dual imperatives of action research. *Information Technology & People*. 14. 46-59. 10.1108/09593840110384771.
14. Pedersen, J.H.H., Sørensen, M. (2020) *Asynchronous Remote Usability Testing – Development and State of the Art: A literature review*. Department of Computer Science, Aalborg University, Aalborg.
15. Pedersen, J.H.H., Sørensen, M. (2020) *Asynchronous Remote Usability Testing in Practice: Exploratory Interviews with IT Professionals*. Department of Computer Science, Aalborg University, Aalborg.
16. Thompson, K. E., Rozanski, E. P., & Haake, A. R. (2004). Here, there, anywhere: Remote usability testing that works. *Sigite 2004 conference* (pp. 132-137). New York and NY: ACM Press.
17. Reeves, S. (2019). How UX Practitioners Produce Findings in Usability Testing. *ACM Transactions on Computer-Human Interaction*, 26, 1–38. https://doi.org/10.1145/3299096
18. Wixon, D., & Dennis. (2003). Evaluating usability methods: why the current literature fails the practitioner. *Interactions*, 10, 28-. https://doi.org/10.1145/838830.838870
19. Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, *DIS, 2006*, 209–218.
20. Hartson, H. R., Castillo, J. C., Kelso, J. and Neale, W. C. Remote evaluation: The network as an extension of the usability laboratory. *Proceedings of CHI 1996*, ACM Press (1996), 228-235.
21. Castillo, J. C., Hartson, H. R. and Hix, D. Remote usability evaluation: Can users report their own critical incidents? *Proceedings of CHI 1998*, ACM Press (1998), 253-254.
22. Hammontree, M. L., Weiler, P., & Nayak, N. P. (1994). Remote usability testing. *Interactions*, 1 , 21-25.
23. Varga, E. (2011) *An Experiential comparative analysis of two remote usability testing methods*. Thesis. Rochester Institute of Technology.

24. McFadden, E., Hager, D.R., Elie, C.J., and Blackwell, J.M. (2002) Remote usability evaluation: Overview and case studies. *International Journal of Human-Computer Interaction* 14(3 and 4), 489–502.
25. Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J., and Uebelbacher, A. (2019) Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing. *Applied Ergonomics* 74, 85–96.
26. Susman, G. and Evered, R. (1978) An Assessment of the Scientific Merits of Action research. *Administrative Science Quarterly* 23, 582-603.
27. Medlock, M. C., Wizon, D. Terrano, M. Romero, R. & Fulton, B. (2002), Using the RITE method to improve products: A definition and a case study. *Proceedings of UPA 2002*. Orlando: Usability Professionals Association,