Evaluating the Usability of a Mobile Guide: The Influence of Location, Participants and Resources

Jesper Kjeldskov[1,2], Connor Graham[1], Sonja Pedell[1],
Frank Vetere[1], Steve Howard[1], Sandrine Balbo[1] and Jessica Davies[3]

[1] Department of Information Systems
The University of Melbourne
Parkville, Victoria 3010
Australia

[2] Department of Computer Science
Aalborg University
DK-9220 Aalborg East
Denmark

[3] Department of Geomatics
The University of Melbourne
Parkville, Victoria 3010
Australia

jesper@cs.auc.dk, cgraham@unimelb.edu.au, spedell@pgrad.unimelb.edu.au,
{showard, fv, sandrine}@unimelb.edu.au, jcdavies@ihug.com.au

ABSTRACT

When designing a usability evaluation, choices must be made regarding methods and techniques for data collection and analysis. Mobile guides raise new concerns and challenges to established usability evaluation approaches. Not only are they typically closely related to objects and activities in the user's immediate surroundings, they are often used while the user is ambulating. This paper presents results from an extensive, multi-method evaluation of a mobile guide designed to support the use of public transport in Melbourne, Australia. In evaluating the guide, we applied four different techniques; field-evaluation, laboratory evaluation, heuristic walkthrough and rapid reflection. This paper describes these four approaches and their respective outcomes, and discusses their relative strengths and weaknesses for evaluating the usability of mobile guides.

INTRODUCTION

Mobile guides constitute a special class of mobile computer system. Usually mobile guides are closely related to the user's physical location and objects in the user's immediate surroundings (e.g. Cheverst et al. 2000, Chincholle et al. 2002, Schmidt-Belz et al. 2002, Reid 2002, Umlauft et al. 2003). Also, they are often used while the user is ambulating, moving from one physical location to another. These properties make the design and evaluation of mobile guides challenging for human-computer interaction researchers and practitioners.

The design of mobile guides has received considerable attention over the last decade (see e.g. Abowd et al. 1996, Cheverst et al. 2000, Cheverst et al. 2002, Pospischil et al. 2002, Fithian et al. 2003). When authors consider the design of mobile guides, they also frequently report the results of evaluations. The reported usability evaluations involve the use of a wide range of methods and techniques borrowed from usability research into 'desk bound' computers and their use, then adapted to fit the special needs, opportunities and limitations of mobile guides. This includes, for example, formal and informal product presentations combined with questionnaires, expert evaluations (Andrade et al. 2002, Po et al. 2004), controlled laboratory experiments (Bohnenberger et al. 2002, Chincholle et al. 2002, Iacucci et al. 2004) and a variety of use studies in realistic field settings including direct observation of use (Cheverst et al. 2002, Schmidt-Belz and Poslad 2003, Laakso et al. 2003), indirect observation of use (Bornträger et al. 2003), field questionnaires (Rocchi et al. 2003), and longitudinal use studies combined with interviews (Kolari and Virtanen 2003, Iacucci et al. 2004). These evaluations all provide valuable insight into usability and usefulness and typically inform design refinements and/or inspire new design concepts. Such research will, one hopes, result in the development of more useful and usable mobile guides.

However, even though evaluations of mobile guides are prevalent, little research has been published on the particular challenges, to usability evaluation, posed by mobile guides; how should we evaluate mobile guides, what methodological challenges do we face, what are the pros and cons of different usability evaluation approaches? Exceptions include, for example, Bornträger and Cheverst (2003) who consider social and technical problems encountered during field evaluations of mobile guide systems, and Kray and Baus (2003) who review and compare nine mobile guide systems and touch upon the methods and techniques that were used in their evaluation. Examining the general literature on mobile HCI does not provide much additional support, with only a few authors considering different usability evaluation methods and techniques for mobile computer systems (see e.g. Brewster 2002, Pirhonen et al. 2002, Kjeldskov and Skov 2003, Kjeldskov and Stage 2004). As a result of our reluctance to 'evaluate evaluation', that is to understand how the utility of the techniques in our usability toolkit respond to the challenge of mobile guide evaluation, no agreed upon set of

usability evaluation methods and data collection techniques exist for mobile guides and little knowledge exists as to when and why one should choose one technique over another. Consequently, researchers and practitioners are provided with little support in making informed decisions about which methods and techniques to select and combine for mobile guide evaluation.

In this paper we report the evaluation of a mobile guide, following four different approaches: field-evaluation, laboratory evaluation, heuristic walkthrough and rapid reflection. The paper describes these four approaches, presents their respective outcomes and discusses their relative strengths and weaknesses from the perspective of the challenges of mobile guide evaluation.

In the next section we present and discuss related research on evaluating the usability of mobile computer systems emphasising the special challenges related to the evaluation of mobile guides. Then we briefly describe a project in which we designed and implemented a mobile guide and evaluated it through four independent usability studies. Each of these usability studies are described in detail, followed by a comparison and discussion of the findings. Finally, we conclude with a number of recommendations for usability evaluation of mobile guides and present avenues for further research.

CHOOSING APPROPRIATE EVALUATION TECHNIQUES

Usability evaluation has proven to be an invaluable tool for ensuring the quality of computerised systems. Usability evaluation of stationary computer systems is an established discipline within human-computer interaction with widely acknowledged techniques and methods (e.g. Dumas and Reddish 1999, Nielsen 1993, Rubin 1994). This is complemented by a growing number of attempts to 'evaluate evaluation', empirical evaluations of the relative strengths and weaknesses of the different approaches and techniques, under different circumstances (e.g. Bailey et al. 1992, Henderson et al. 1995, Karat et al. 1992, Molich et al. 1998). So far, this kind of research is only beginning to emerge in relation to the evaluation of mobile computer systems.

Mobile guides take many of the well known methodological challenges of evaluating the usability of both stationary and mobile computer systems to an extreme. Users of mobile guides are ambulatory, typically highly mobile during their interaction with the system, and are situated in a dynamic and often unknown use setting (e.g. Tamminen et al. 2003, Vetere et al. 2003, Schmidt-Belz et al. 2002, Makimoto and Manners 1997). Furthermore, the information presented to the users of mobile guides is closely related or indexed to their physical location, objects in their immediate surroundings and to their present as well as planned activities (e.g. Chincholle et al. 2002, Pospischil et al. 2002, Kolari et al. 2003, Kray and Baus 2003, Kjeldskov et al. 2003). The questions and challenges related to choosing appropriate techniques for evaluating the usability of mobile guides are several. Should the evaluation be done in the lab or in the field? Should the evaluation be based on usability experts and/or involve users? How should the data be analyzed; using a thorough (but time consuming) qualitative and quantitative analysis or a 'discount' approach?

In-Situ or In-Vitro?

Since the use of mobile guides is so closely related to the user's context, evaluating in the field seems like an appealing, even indispensable, approach. Indeed most existing studies of mobile usability apply some type of field-based approach. Yet, as the relative strengths and weaknesses of laboratory and field-based methods and techniques for evaluating mobile devices become better understood, this assumption is challengeable (Kjeldskov et al. 2004, Po et al. 2004). Applying a laboratory-based approach, evaluations can benefit from experimental control and high quality data collection. Yet traditional usability laboratory setups may not adequately simulate the context surrounding the use of mobile systems. Using a field-based approach, it may be possible to obtain a higher level of 'realism'. However, field-based usability evaluations are not easy (Brewster 2002, Nielsen 1998) and applying established evaluation techniques and data collection instrumentation, such as multi-camera video recording, think-aloud protocols or shadowing may be difficult in natural settings (Sawhney and Schmandt 2000). Also, field evaluations complicate data collection since users are moving physically in an environment over which we have little control (Johnson 1998, Petrie et al. 1998) and only partially comprehend.

Users, Surrogates or Experts?

Usability evaluations in both laboratories and in-situ are problematic for mobile technology because they involve techniques that assume usage that is relatively fixed, tasks that endure over a reasonable period of time and (for laboratory evaluations) can be de-contextualised easily. Furthermore, laboratory and field based evaluations typically involve studying prospective user's interaction with the system being evaluated. This can be very time consuming and hampered by limited access to participant's unfamiliar with the process. As an alternative, usability research has promoted a tranche of expert-based evaluation techniques, such as heuristic inspection (Nielsen and Molich 1990) and cognitive walkthrough (Wharton et al. 1994) which may offer benefits. These techniques typically benefit from providing evaluators with guidance (in the form of heuristics or a checklist) for identifying a prioritised list of usability flaws. However, inspection approaches are often criticised for finding proportionately fewer problems in total, and disproportionately more cosmetic problems (Karat et al. 1992). Further, inspection based approaches have been accused of *context immunity* (Po et al. 2004).

Exhaustive or Discount Data Analysis?

One of the most resource demanding activities in a usability evaluation is the analysis of collected empirical data, a stage vital to lessons learned, and yet difficult and time consuming to conduct. Whereas there is a strong body of research within human-computer interaction regarding the appropriate choices of data collection methods and techniques, data analysis is vaguely described by many authors, e.g. (Nielsen 1993, Preece et al. 1994, Rubin 1994). Many methods and techniques exists for analyzing the empirical data from usability evaluations like, for example, grounded analysis (Strauss and Corbin 1998), video data analysis (Nayak et al. 1995, Sanderson and Fisher 1994), cued-recall (Omodei et al. 2002), and expert analysis (Nielsen and Molich 1990), etc. However, approaches to instrumenting data analysis are often poorly discussed (Gray and Saltzman 1998) and the relative value of applying such exhaustive approaches to the analysis of usability data is still largely speculative. Of special note, it seems implicitly assumed by many authors that a thorough grounded analysis or video analysis with detailed log-files and transcriptions of usability evaluation sessions is the gold standard by which evaluation should be judged (Sanderson and Fisher 1994). However, the balance between the costs of spending large amounts of time on video analysis and the value added to the subsequent results has been questioned (Nielsen 1994) and is an open question in relation to the evaluation of mobile guides.

THE TRAMMATE PROJECT

Inspired by the challenges discussed above, during 2002 and 2003 we explored the issues surrounding the design and evaluation of a mobile guide.

We conducted a research project focusing on the potential of mobile guides for supporting the use of public transportation in Melbourne, Australia (Kjeldskov et al. 2003). The project was motivated by discussions among consultants and sales staff of a large IT company regarding alternatives to the use of cars for traveling in the city to meetings with clients. In large cities where traffic is often very dense, traveling by car can be time-consuming, necessitating much planning. Using Melbourne's tram-based public transport would not only have environmental benefits, but might also be more effective if supported by a mobile information service providing travelers with relevant information at the right time and in the right place.

From this study, we identified some key requirements for a mobile guide supporting the use of the public transportation system:

- Relating travel information directly to the users' unfolding schedule of formal and informal appointments;
- Providing route-planning information for the tram system based on the user's current location and time;
- Alerting the users when it is time to commence their journey in order to make it to the destination in time;
- Providing easy access to key information such as travel time, walking distance and number of route changes.

The Prototype System

A functional mobile guide prototype for Melbourne's tram system was developed by researchers at the University of Melbourne's Department of Geomatics (Smith et al. 2004). The prototype provided route-planning facilities for the tram system based on the user's current location as a combination of textual instructions and annotated maps, satisfying some of the requirements described above. One of the overall screens in the prototype system is shown in figure 1.

Figure 1. Entering a destination into the mobile guide

The prototype was designed for an iPAQ handheld computer equipped with a WAP browser. The device is connected to the Internet via a GPRS data connection and acquires its position via GPS. The application was designed to serve three functional processes with regard to public transport. These were accessible via the startup screen.

1. Timetable Lookup: information about the tram timetable based on the input of stop numbers (origin and destination) and route numbers. This function was aimed at regular tram users who are very familiar with their route of travel. No maps are available within this section of the system.
2. Plan Trip: information about the whole route (containing route descriptions and maps) based on the input of suburb and street corners of origin and desired destination. Users were also presented with an option to enter an arrival time or departure time for their journey. From each screen within this function, it was possible to view a visual representation of the relevant portion of the journey on a map.

3. Determine Route: information about the whole route (containing route descriptions and maps) based on the input of the street corner of the destination and the suburb. The system determined the user's origin location via a GPS. Maps were also available for components of the journey in this function.

Upon entering all required input, the system computes a suitable travel plan for using the tram network between the desired origin and destination. The solution suggested by the system is optimal in terms of normative data on journey length (measured in number of stops), and the timing of tram vehicles. An example of the maps displayed by the system is shown in figure 2.

Figure 2. Map view on the mobile guide

COMPARING THE FOUR APPROACHES

In order to investigate the advantages and disadvantages of different techniques for evaluating the usability of mobile guides, we conducted four different evaluation studies of the mobile guide prototype described above:

1. Field Evaluation: exhaustive analysis of user-based data; data collected in-situ but analysed in-vitro
2. Laboratory Evaluation: exhaustive analysis of user based data; data collection and analysis conducted in-vitro
3. Heuristic Walkthrough: discount collection and analysis of usability problems by experts; data collection and analysis conducted in-vitro
4. Rapid Reflection: discount analysis of user-based data from field and laboratory studies; in-vitro data analysis. This analysis was done prior to the exhaustive analysis in studies 1 and 2

These four evaluations are described in detail in the following sections.

Study 1: Field Evaluation

The field evaluation focused on guide use in realistic settings. It took place over two days in the city centre of Melbourne, Australia. The evaluation involved five test subjects between twenty one and forty two years of age. The test subjects were all frequent computer users and had experience with the use of PDAs and mobile phones. The test subjects were all familiar with the tram system of Melbourne.

Figure 3. Field evaluation of the mobile guide

The subjects had to complete four realistic tasks involving route planning while traveling to appointments in the city by tram. The tasks were derived from the earlier user studies in the TramMate project and were piloted prior to the evaluation, resulting in minor modifications in order make them achievable within a feasible timeframe. In order to solve the tasks, the test subjects had to lookup information available in the mobile guide and then perform the tasks 'for real' (e.g. catching a tram to a specific destination). An example task is shown below:

> You are going to catch a tram from the corner of Swanston and Queensberry Street in Carlton for a meeting at the corner of Little Collins and Exhibition Street in Melbourne. You have to be there in about 30 minutes from now.
>
> Using the plan trip option, find out:
>
> **a.** Which tram route(s) to take
>
> **b.** When the first possible tram is departing
>
> **c.** The number of route changes (if any)
>
> **d.** If there is a route change, where to board the second tram.
>
> **e.** Which stop to get off the last tram.
>
> **f.** How to get from the last stop to your final destination.
>
> **g.** The estimated time of arrival.
>
> Use this information to get to the meeting.

The prototype accessed live timetable information through a GPRS connection to the Internet. Due to technical problems with acquiring precise GPS positioning data in the city area and on the trams, positioning was simulated by the researchers by inputting predefined spatial data into the system 'behind the scenes' of the evaluation. Users were not aware of this.

The field evaluation involved four people for each evaluation session. One test subject used the mobile guide to solve the tasks. One researcher managed the evaluation sessions, encouraging the test subjects to think-aloud and asking questions for clarification similar to a contextual interview. Another researcher recorded the evaluation sessions on video switching between close-up views of the device and overall views of the surroundings. A third researcher took written notes (figure 3).

The data from the field evaluation was subjected to a detailed grounded analysis (Strauss and Corbin 1998), producing a list of richly described usability problems. The problems were rated as critical, serious or cosmetic in accordance with Molich (2000).

Critical problem

- Recurred across all users
- Stopped users completing tasks

Serious problem

- Recurred frequently across users
- Inhibited /slowed down users completing tasks
- Users could (eventually) complete tasks

Cosmetic problem

- Did not recur frequently across users
- Did not inhibit users severely
- Users could complete tasks

The time spent on the field evaluation amounted to fifty six person-hours for data collection and twenty six person-hours for data analysis.

Study 2: Laboratory Evaluation

The laboratory evaluation focused on use in a controlled setting. It was conducted in a state-of-the-art usability laboratory at the University of Melbourne's Department of Information Systems. Due to less time required for logistics, we were able to conduct the laboratory evaluation in one day.

We intentionally designed the laboratory evaluation to be similar to the field evaluation in a number of important ways as this allowed us to compare the results across techniques. However some differences were necessary if we were to 'play to the strengths' of each approach. The laboratory evaluation involved the same number and type of test subjects (between twenty one and twenty five years of age) and the test subjects had to solve the same four tasks using the same mobile guide system. However, in the laboratory evaluation, the subjects were seated at a desk, with the mobile guide in their hand rather than being physically mobile. Also, they did not have to perform the tasks 'for real' as in the field, that is they were not required to board a tram and take the journey.

The laboratory setting allowed for high-quality audio and video recordings from multiple perspectives (figure 4). Three ceiling-mounted cameras captured overall views of the test subject and test monitor. A fourth camera on a tripod captured a close-up view of the mobile guide (figure 5). To ensure a good view of the screen and interaction, the test subjects were asked to hold the device within a limited physical area indicated on the table.

Figure 4. Laboratory evaluation of the mobile guide

As in the field, the mobile guide accessed live timetable information while positioning was simulated. The laboratory evaluation involved four people: one test subject and three researchers; a test monitor or host, encouraging the test subject to think aloud and asking questions for clarification; and two data loggers, observing the evaluation through a one-way mirror respectively. The data from the laboratory evaluation was analyzed using the same method as for the field evaluation, resulting in a similar list of identified usability problems.

Figure 5. Close-up of interaction with the mobile guide

The time spent on the laboratory evaluation amounted to thirty two person-hours for data collection and eighteen person-hours for data analysis.

Study 3: Heuristic Walkthrough

The third evaluation of the mobile guide focused on usability as perceived by experts in human-computer interaction. It was conducted in the same laboratory used for the laboratory study (figure 6) and consisted of a heuristic walkthrough guided by a set of heuristics developed specifically for the purpose of this evaluation,

heuristics sensitive to the mobile challenge. For a detailed description see Vetere et al. (2003).

Four evaluators, all with expertise in HCI and usability, each independently performed a heuristic walkthrough of the mobile guide. The evaluators were given the mobile guide heuristics and a common set of tasks to contextualize the evaluation, thereby blending aspects of traditional heuristic evaluation and the cognitive walkthrough. The tasks were the same as used in the field and laboratory evaluations.

Each evaluation lasted an average of one and one quarter hours. First, the evaluators were welcomed by the host (a representative from the design team), and given the opportunity to ask questions about the process. The evaluators then explored the device, without reference to either the heuristics or the task scenarios. Thereafter, the evaluators assessed the device against the heuristics and recorded their observations. Finally, the evaluators worked through each task, recording further observations against the heuristics. After all heuristic walkthroughs had been completed results were collated in a post session workshop, allowing the evaluators to discuss their identified usability problems. As in the field and laboratory evaluations, the mobile guide accessed live timetable information, while positioning was simulated.

Figure 6. Heuristic walkthrough

All but one of the evaluators completed all tasks, and all evaluators addressed the mobile guide heuristics. Additionally all evaluators drew broadly on their knowledge of usability, not confining themselves to 'mobility issues' or the mobile guide heuristics alone, and all reflected on the heuristic walkthrough process itself.

The time spent on the heuristic walkthrough amounted to ten person-hours in total.

Study 4: Rapid Reflection

The fourth study had the purpose of investigating the potential for reducing the effort spent on data analysis by applying a 'rapid reflection' approach inspired by rapid ethnography (Millen 2000). The rapid reflection study of the mobile guide differed somewhat from the other three studies. Rather than being a completely separate study, the rapid reflection approach was based on the empirical data gathered through the field and laboratory evaluations. However, as an alternative to the rather time consuming grounded analysis of the video data, the rapid reflection approach applied a pragmatic discussion and consideration of the collected data by the involved evaluators. For a detailed description of this study see Pedell et al. (2003).

The rapid reflection sessions (figure 7) followed immediately after the field and laboratory evaluations and involved all participating researchers. On the basis of the observers' written notes and printed screen shots from the evaluated mobile guide, the rapid reflection sessions had the purpose of discussing and agreeing upon what main themes and usability problems that emerged on that specific day. Each session was time-boxed at one hour.

Figure 7. Rapid Reflection session

The rapid reflection session was assisted by an observer, who was not present during the laboratory or field evaluations, asking questions for clarification. Furthermore, one of the researchers had the role of writing all identified usability problems and other issues on a whiteboard as they were presented, and keeping an overview of the discussed usability problems as the session progressed. After the reflection session, one of the researchers spent another hour on writing up the contents of the whiteboard into a richly described list of usability issues, which was then circulated among the researchers for validation and comments.

The time spent on the rapid reflection approach amounted to a total of fourteen person-hours for the field data and eight person-hours for the laboratory data. As the rapid reflection builds on the data already collected in the field (study 1) and lab (study 2) respectively, these numbers should be compared to the twenty six and eighteen hours spent on the exhaustive data analysis described above.

Analysis

The analysis of data from each of the four approaches described above focused on identifying and describing usability problems experienced with the use of the mobile guide prototype. In the case of the field and laboratory evaluation this was done through the use of grounded analysis (Strauss and Corbin, 1998). In the case of the heuristic walkthrough and the rapid reflection it was done through post-evaluation workshops. Two discrete steps were involved in the comparison of the results across the

four approaches; a *compilation* of the results and a *comparison* of the results across techniques. In order to ensure that this process was rigorous and that both the compilation and comparison of results were credible, dependable and confirmable (Lincoln and Guba 1986) the following steps were taken.

Firstly, one researcher compiled the results for each of the four approaches into four lists of identified usability issues. This researcher was involved in data analysis for the field and laboratory evaluation, and data collection and analysis for the rapid reflection and heuristic walkthrough. Thus this researcher had proximity to the results from each of the four approaches, a prolonged engagement with the results and had engaged in persistent observation of the data (Guba and Lincoln 1989). Following the compilation of the results from the four different approaches, all participating researchers were required to revisit the list from each approach. In this way, the dependability (Guba and Lincoln 1989: 242) of the results for each of the four approaches was ensured. Secondly, another researcher (who had been involved in the data collection for the field and laboratory evaluation and rapid reflection) collaborated with the first researcher in the compilation of the results for each of the four approaches into one merged list. This collaboration involved extended discussions of the identified problems (member checking) and in the monitoring of the compilation of the results for each of the four approaches (progressive subjectivity) (Guba and Lincoln 1989). In case of different severity ratings of the same usability issue across techniques, the most severe rating was used in the merged list. To be able to identify disparities in severity ratings the original ratings were preserved as comments to each of the cells in the list. Finally, the merged list of usability issues was presented and discussed jointly by the full team of participating researchers (the authors of this paper) through a one-hour workshop. This was done to ensure that the comparisons across techniques were credible (through member checking and the involvement of the attendant researchers in the initial analysis), and dependable and confirmable (through an audit of the results and comparisons by two researchers). The resulting list of merged problems can be found in the appendix.

In the next section we present our findings, and draw out some key differences between the four approaches as they apply to the task of evaluating a mobile guide. Differences between the approaches that are not germane to mobile guide evaluation are outside the scope of this paper.

It should be noted that, in presenting our results, we do not claim statistical power, but rather aim to present a rich, qualitative overview of the data, drawing out differences and similarities as they arise. This allows us to draw some overall conclusions concerning the pros and cons of different techniques for evaluating the usability of mobile guides.

FINDINGS

Jointly, the four usability studies generated a list of twenty two distinct usability problems. Of these twenty two problems, a total of five problems were classified as critical, eleven as serious, and six as cosmetic (see final column of table 1). Critical usability problems related to the interaction between the user/system and the surrounding environment, for instance the representation of map and textual information in the system and the way the system required the user to use this information. Another critical issue was caused by disparities in the relationship between information presented in the system and the context in which the user was situated. Critical problems were typically related to mapping issues arising from the use of the 'system in the world'.

The distribution of usability problems across the four approaches is summarized in table 1.

	Field evaluation	Lab evaluation	Heuristic walkth.	Rapid Reflection	Total
Critical	4	4	4	4	5
Serious	7	6	6	5	11
Cosmetic	2	3	3	4	6
Total	13	13	13	13	22

Table 1. Distribution of the number of usability problems identified using the four different techniques

Regarding problem coverage, any individual technique identified little more than half of the total problem set (coincidently, thirteen from twenty two in each case).

Looking at the critical problems, all techniques identified four out of five critical problems, no technique identifying all problems. In the case of serious problems, more variation was observed across the four techniques, with the identification of between five and seven problems, from a total set of eleven. Again, no single technique was able to identify all eleven issues, and only the field evaluation identified more than half of the total number of serious problems. In the case of cosmetic problems, the rapid reflection technique was the most effective, identifying four out of six problems. While missing two of the five cosmetic problems identified through the video analysis, the rapid reflection was the only technique that reported the issue of problems with using the system causing strong emotional responses from the users. As an interesting aside, it should be noted that the heuristic walkthrough did not generate the usual level of 'cosmetic noise' that often characterizes expert evaluations based on general usability heuristics (Karat et al. 1992). It may be that tailoring the heuristics (see Vetere et al. 2003) to the mobile problem helped reduce such noise, especially false positives, in the data.

The distribution of problems identified across the four techniques is illustrated in figure 8.

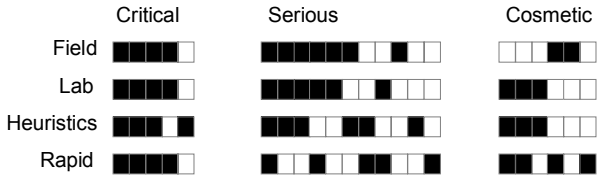

Figure 8. Distribution of usability problems. A black square indicates that a problem was identified using that specific technique. A white square indicates that a problem was not identified using that specific technique but was found using another technique.

Figure 8 shows twenty two usability problems (each column represents a specific problem), stratified as critical, serious or cosmetic, distributed across the four different techniques. A black square shows that a problem was identified using that technique. A white square indicates that a problem was not identified using that technique, but was found using another technique (see appendix 1 for a brief description of the problems).

The distribution of problems in figure 8 is discussed below.

Critical Problems

Three out of the total set of five critical problems were identified by all techniques, with a further problem identified by all but the heuristic walkthrough. Though comparing evaluation approaches is always challenging, due primarily to the lack of any independently established problem set, we can be confident that these four critical problems were indeed present in the evaluated mobile guide, rather than being false positives. On the other hand, the distribution of critical problems also indicates that the identification of critical problems depended little on the precise circumstances surrounding the deployment of a specific evaluation approach; it is encouraging that critical problems generally are uncovered regardless of approach. It is also noticeable that the field, lab and rapid reflection studies were consistent in the types of critical problems identified.

For the identification of the most severe issues in a mobile guide, discount data analysis appears to be adequate. The benefits of an exhaustive grounded analysis may not out way the associated costs.

Only one critical usability problem was unique to a specific approach. This 'problem', identified by the heuristic walkthrough, concerned the general purpose of the guide, and its alignment with broader lifestyle and use issues not evident in findings drawn from the other approaches. Issues raised here included the degree to which users could flexibly adapt the device to fit lifestyle activity (Vetere et al. 2003).

The critical problem not identified in the heuristic walkthrough was a problem related to disparities in the relationship between information in the system, and the users' context- the 'system in the world' problem referred to earlier. This problem was adjudged critical in both the field and rapid reflection studies (which in turn drew on the data collected in the field), but cosmetic in the laboratory study. Given the situated flavour of this problem, the different severity ratings are not surprising. However, it does highlight the fact that while contextually related problems may appear in laboratory settings, they can be experienced, and described, in very different ways compared to the field.

Serious Problems

The distribution of serious problems shows a more varied picture across approaches. Of eleven serious problems, eight were identified by two or more of the techniques, four were found by three techniques or more, and only one problem was identified by all techniques. Three serious problems were uniquely identified by only one technique.

Whereas the critical problems reflected 'system in the world' issues, serious problems were more oriented to significant usability hurdles: difficulty in entering data into the system, difficulty in being able to recover from errors and poor labelling of interface elements. Additionally, the systems' implicit assumptions about the users' existing knowledge of the city in which the mobile guide was used also drew attention here. Other serious problems related to cognitive load demands, e.g. remembering data from one screen when interacting with another, and lack of flexibility to deviate from a predefined, by the system, path of interaction.

Looking at the clustering of problems, it is noticeable that there is a relatively large overlap between the findings from the field and laboratory studies. Five out of the total eleven serious problems were identified in both the lab and the field, with the field identifying only two additional unique problems and the laboratory only one further unique problem. The five serious problems identified in both the laboratory and the field included the four most prominent; input, recovery and labelling.

Whilst some of the more serious flaws were also identified by both the heuristic walkthrough and the rapid reflection, and both of these approaches contributed unique problems (one in each case), both the heuristic walkthrough and the rapid reflection missed four and five serious problems respectively, from those identified collectively in the field and in the lab.

Cosmetic Problems

The picture is yet more confused when examining cosmetic problems. None of the cosmetic problems were identified by all techniques, and only two problems were identified by three of four approaches.

Looking at the clustering of problems, there was no overlap between the cosmetic problems found in the field and in the lab. The field approach drew attention to issues such as the

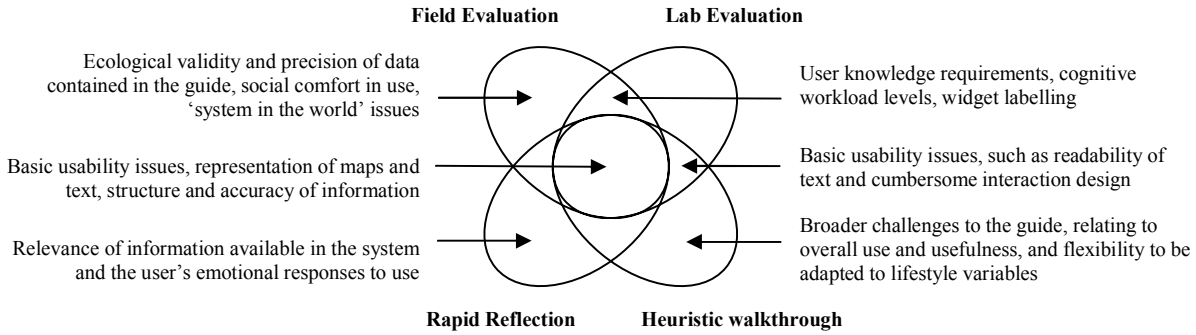

Figure 9. Schematic overview of the types of usability issues identified in overlaps between the different techniques

real-world validity and precision of the data presented by the system and the 'social comfort' (e.g. whether it felt embarrassing to use the device in a public setting). In contrast, the lab based approach drew attention to device oriented issues, such as the readability of text and efficiency of looking up information.

Interestingly, the lab and the heuristic walkthrough identified the same problem set, with the rapid reflection sitting somewhere in-between, identifying one unique problem related to the observation, that many users had a strong emotional response when encountering problems with the system.

In the next section we draw out general lessons learned, especially in relation to the similarities and differences between the four approaches.

DISCUSSION

Figure 9 outlines the overlap between the four approaches, in terms of the usability problems identified.

There are benefits to be gained from each approach in relation to the types of usability problems uncovered, but many strengths are shared by more than one technique. The cluster in the centre of figure 9 emphasises that many usability issues related to the representation, accuracy and structure of the map and textual information provided, and these issues are captured by all approaches.

All approaches, with the exception of the laboratory study, identified unique problems. The field evaluation uniquely identified issues of validity and precision of the data presented by the device, and the lack of social comfort when using the device in public. The heuristic walkthrough uniquely identified issues related to the overall use and usefulness of the mobile guide, and its flexibility in relation to different user activities. The rapid reflection approach, though based on the data from the lab and field studies, brought forward some issues related to the perceived relevance of available information and highlighted the users' strong emotional responses (ranging from frustration to sheer outrage!) to the hurdles presented by the design.

Examining the various pair-wise comparisons, it is interesting to note that the overlap between the laboratory evaluation and heuristic walkthrough contains basic usability problems, such as the readability of screen text, whereas the overlap between the field and laboratory studies contains the potentially more complex problems of the assumed extent of users' prior knowledge and the cognitive workload demands placed on the user.

Contrasting the laboratory and field studies, two differences in the problem sets are worthy of note. Whilst the laboratory problems were reported in great detail (often related to the artefact per se, for example, mislabelling of commands), the field study stressed problems of mobile 'use' rather than simply device usability, and typically those problems were expressed in the language of the situation. For example, spending too long inputting commands was made urgent through making explicit the pressing demands of the situation; the user might be stationary, reading the mobile display, and blocking a footpath in the situation of use.

The rapid reflection sessions briefly summarized the key issues from the field and laboratory user studies requiring considerably less person-hours for analysis. Generally, the problems reported through the rapid reflection were less specific and the list of problems was not complete compared to the joint outcome from the video analysis. On the other hand, the rapid reflection technique allowed the researchers to focus only on the top-most severe problems observed. Identifying four out of five critical problems in less than half the time required for the video analysis, the rapid reflection proved to be a very cost-effective usability analysis technique. This finding is consistent with a similar comparison done by Kjeldskov et al. (2004).

Across the four approaches there is much similarity in the pictures that emerge of the mobile guide, but there are

many compelling differences. We will now summarise some general lessons learned.

In-Situ or In-Vitro?

The development of electronic mobile guides remains a rather recent design challenge, and we cannot rely on established theory or rigorously tested examples of best practice to guide us. Collecting data in-situ prompted us with elements of the situation of use that we might have been ignorant of, or that might have passed un-remarked. Additionally, being in-situ provoked a very concrete consideration of how things might be changed; it is easy to be lazy when discussing the future, speculations turning from plausible fiction to science fiction. Being in-situ was our insurance policy against ignorance in the absence of a refined understanding of what 'the situation of use' was, or might become.

Until we are able to supplement our meagre understanding of mobile use, and unless there are insurmountable practical or logistical hurdles to accessing the situation of use, we should continue to collect, at least as a part of a broader data collection protocol, data in the field.

Users, Surrogates or Experts?

The issue of expert versus user based evaluation is part of a more general discourse (for example Dumas and Redish 1999, Nielsen 1994) that we will not cover here. In respect of mobile guides, a few comments are appropriate.

Due to the relative novelty of mobile guides, and the lack of a substantial relevant knowledge base, the perceived 'opinion free' flavour of user based tests, as compared to inspection based approaches, might strengthen the usability argument in the broader software development process. In contrast, the relative novelty of the mobile guide paradigm should drive us to 'test early and often'; anecdotally, experts are able to overcome the credibility hurdles involved in early paper-based prototypes more ably than end-users.

Exhaustive or Discount?

Our activities in the development of mobile guides are thirsty for foundational concepts and theoretical insight. The motivation for exhaustive data collection and analysis extends beyond theory building, to practice as it relates to safety critical or business critical applications. We should continue to champion discount approaches for the fast cycle, discovery oriented phases of early product development, whilst encouraging a concerted effort in building the theoretical foundations of an applied science of mobile use.

CONCLUDING COMMENTS

Whilst no individual approach to the usability testing and evaluation of mobile guides can be held to be the definitive approach, any testing and evaluation is much better than none at all. The level of agreement amongst the approaches was both significant and encouraging, but not complete and multi-method approaches to mobile guide evaluation are clearly useful, as implied in figure 9.

Mobile guides raise particular if not unique challenges, including the need to understand the users' experience of the 'system in the world', establishing and designing for social comfort and the evaluating the compatibility between the device and broader lifestyle considerations. These particular challenges provide new reasons to respect the unfolding nature of, and situated character of, the interactions between people and technology. New challenges that, with time, will be met by advances in our theoretical apparatus, our methodological toolkit, and our sense of what is and what is not best practice in relation to the design of mobile guides.

ACKNOWLEDGMENTS

Thanks to the consultants who acted as participants in the initial user studies, thanks to the lab and field test subjects, and thanks to those colleagues involved in the earlier phases of the TramMate project: Jennie Carroll, John Murphy, Jeni Paay and Daniel Tobin. This research was supported by the Danish Technical Research Council (project reference 26-03-0341).

REFERENCES

ABOWD, D., ATKESON, C., HONG, J., LONG, S. and PINKERTON, M., 1996, Cyberguide: A Mobile Context-Aware Tour Guide. *Wireless Networks,* **3(5),** 421-433.

ANDRADE, M. T., SANTOS, E., LIVADITI, J., and TSAKALI, M., 2002, Managing multimedia content and delivering services across multiple client platforms using XML. Proceedings of Mobile Tourism Support (Pisa, Italy: in conjunction with Mobile HCI 2002).

BAILEY, R. W., ALLAN, R. W., and RAIELLO, P., 1992, Usability Testing vs. Heuristic Evaluation: A Head-to-Head Comparison. Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting, HFES, pp. 409-413.

BOHNENBERGER, T., JAMESON, A., KRÜGER, A. and BUTZ, A., 2002, Location-Aware Shopping Assistance: Evaluation of a Decision-Theoretic Approach. Proceedings of Mobile HCI 2002 (Pisa, Italy: LNCS, Springer-Verlag).

BORNTRÄGER, C., CHEVERST, K., DAVIES, N., DIX, A., FRIDAY, A. and SEITZ, J., 2003, Experiments with Multi-modal Interfaces in a Context-Aware City Guide. Proceedings of Mobile HCI 2003 (Udine, Italy: LNCS, Springer-Verlag), pp. 116-130.

BORNTRÄGER, C., and CHEVERST, K., 2003, Social and Technical Pitfalls Designing a Tourist Guide System. Proceedings of HCI in Mobile Guides (Udine, Italy: in conjunction with Mobile HCI 2003).

BREWSTER, S., 2002, Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing,* **6**, 188-205.

CHEVERST, K., DAVIES, N., MITCHELL, K., 2002, Exploring Context-Aware Information Push. *Personal and Ubiquitous Computing,* **6**, 276-281.

CHEVERST, K., DAVIES, N., MITCHELL, K., FRIDAY, A. and EFSTRATIOU, C., 2000 Developing a Context-Aware Electronic Tourist Guide: Some Issues and Experiences. Proceedings of CHI'00 (The Hague, Netherlands: ACM), pp. 17-24.

CHINCHOLLE, D., GOLDSTEIN, M., NYBERG, M. and ERIKSON, M., 2002, Lost or Found? A Usability Evaluation of a Mobile Navigation and Location-Based Service. Proceedings of Mobile HCI 2002 (Pisa, Italy: LNCS, Springer-Verlag), pp. 211-224.

DUMAS, J. S., and REDISH, J. C., 1999, *A Practical Guide to Usability* Testing (Exeter: Intellect).

FITHIAN, R., IACHELLO, G., MOGHAZY, J., POUSMAN, Z. and STASKO, J., 2003, The Design and Evaluation of a Mobile Location-Aware Handheld Event Planner. Proceedings of Mobile HCI 2003 (Udine, Italy: LNCS, Springer-Verlag), pp. 145-160.

GRAY, W. D., and SALTZMAN M. C., 1998, Damaged Merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction,* **13(3)**, 203-261.

GUBA, E. G., and LINCOLN, Y. S., 1989, *Fourth Generation Evaluation*. (California: Sage Publications).

HENDERSON, R., PODD, J., SMITH, M., and VARELA-ALVAREZ, H., 1995, An Examination of Four User-Based Software Evaluation Methods. *Interacting with Computers,* **7(4)**, 412-432.

IACUCCI, G., KELA, J. and PEHONEN, P., 2004, Computational support to record and re-experience visits. *Personal and Ubiquitous Computing,* **8(2)**, 100-109.

JOHNSON, P., 1998, Usability and Mobility: Interactions on the move. Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices.

KARAT, C. M., CAMPBELL, R., and FIEGEL, T., 1992, Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. Proceedings of CHI'92 (New York: ACM) pp. 397-404.

KOLARI, J., and VIRTANEN, T., 2003, In the Zone: Views through a context-aware mobile portal. Proceedings of HCI in Mobile Guides (Udine, Italy: in conjunction with Mobile HCI 2003).

KJELDSKOV, J., SKOV, M. B., and STAGE, J., 2004, Instant Data Analysis: Conducting Usability Evaluations in a Day. Proceedings of NordiCHI 2004 (Tampere, Finland: ACM).

KJELDSKOV, J., HOWARD, S., MURPHY, J., CARROLL, J., VETERE, F. and GRAHAM, C., 2003, Designing TramMate - a context aware mobile system supporting use of public transportation. Proceedings of DUX 2003 (San Francisco, CA, USA: ACM).

KJELDSKOV, J., and STAGE, J., 2004, New Techniques for Usability Evaluation of Mobile Systems. *International Journal of Human-Computer Studies,* **60**, 599-620.

KJELDSKOV, J., SKOV, M. B., ALS, B. S., and HØEGH, R. T., 2004, Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. Proceedings of Mobile HCI 2004 (Glasgow, Scotland: LNCS, Springer-Verlag).

KJELDSKOV, J. and SKOV, M. B., 2003, Creating a Realistic Laboratory Setting: A Comparative Study of Three Think-Aloud Usability Evaluations of a Mobile System. Proceedings of Interact 2003 (Zürich, Switzerland: IOS Press).

KRAY, C., and BAUS, J., 2003, A Survey of mobile guides. Proceedings of HCI in Mobile Guides (Udine, Italy: in conjunction with Mobile HCI 2003).

LAAKSO, K., GJESDAL, O., and SULEBAK, J. R., 2003, Tourist information and navigation support by using 3D maps displayed on mobile devices. Proceedings of HCI in Mobile Guides (Udine, Italy: in conjunction with Mobile HCI 2003).

LINCOLN, Y. S., and GUBA, E. G., 1986, But is it rigorous. Trustworthiness and authenticity in naturalistic evaluation. In *Naturalistic Evalution*, edited by D.D. Williams (San Francisco: Jossey-Bass).

MAKIMOTO, T., and MANNERS, D., 1997, *Digital Nomad* (Chichester: John Wiley & Sons).

MOLICH, R., 2000, *Usable Web Design* (In Danish) (Ingeniøren | bøger)

MOLICH, R., BEVAN, N., CURSON, I., BUTLER, S., KINDLUND, E., MILLER, D., and KIRAKOWSKI, J., 1998, Comparative Evaluation of Usability Tests. Proceedings of the Usability Professionals Association Conference, pp. 189-200.

NAYAK, N. P., MRAZEK, D., and SMITH, R.D., 1995, Analysing and Communicating Usability Data. *SIGCHI Bulletin,* **27 (1)**.

NIELSEN, C., 1998, Testing in the Field. Proceedings of the third Asia Pacific Computer Human Interaction Conference, APCHI 1998 (IEEE Computer Society)

NIELSEN, J., 1994, Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. In *Cost-Justifying Usability,* edited by R. G. Bias, and D. J. Mayhew (Academic Press Professional).

NIELSEN, J., 1993, *Usability Engineering.* (Morgan Kaufmann).

NIELSEN, J. and MOLICH, R., 1990, Heuristic evaluation of user interfaces. Proceedings of CHI'90 (Seattle, USA: ACM Press), pp. 249 – 256.

OMODEI, M. M., WEARING, A. J., and MCLENNAN, J., 2002, Head-Mounted Video and Cued Recall: A Minimally Reactive Methodology for Understanding, Detecting and Preventing Error in the Control of Complex Systems. Proceedings of the 21st European Annual Conference of Human Decision Making and Control., (Scotland: University of Glasgow).

PEDELL, S., GRAHAM, C., KJELDSKOV, J. and DAVIES, J., 2003, Mobile Evaluation: What the Metadata and the data told us. Proceedings of OzCHI 2003 (Brisbane, Australia: CHISIG)

PETRIE, H., JOHNSON, V., FURNER, S., and STROTHOTTE, T., 1998, Design Lifecycles and Wearable Computers for Users with Disabilities. Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices (Glasgow).

PIRHONEN, A., BREWSTER, S.A., and HOLGUIN, C., 2002, Gestural and audio metaphors as a means of control for mobile devices. Proceedings of CHI 2002 (ACM).

PO, S., HOWARD, S., VETERE, F., and SKOV, M., 2004, Heuristic Evaluation and Mobile Usability: Bridging the Realism Gap. Proceedings of Mobile HCI 2004 (Glasgow, Scotland: LNCS, Springer-Verlag).

POSPISCHIL, G., UMLAUFT, M. and MICHLMAYR, E., 2002, Designing LoL@, a Mobile Tourist Guide for UMTS. Proceedings of Mobile HCI 2002 (Pisa, Italy: LNCS, Springer-Verlag), pp. 140-154.

PREECE, J., ROGERS, H., SHARP, D., BENYON, D., HOLLAND, S., and CAREY, T., 1994, *Human-Computer Interaction*. (Workingham: Addison-Wesley).

REID, H., 2002, Maps on your PDA - Simple Stuff #8 - Mapping the Trade Show Floor. *Directions Magazine 2002,* http://www.directionsmag.com/article.php?article_id=199 (accessed 2 June 2004)

ROCCHI, C., STOCK, O., and ZANCANARO, M., 2003, Semantic-based Multimedia Representations for the Museum Experience. Proceedings of HCI in Mobile Guides (Udine, Italy: in conjunction with Mobile HCI 2003).

RUBIN, J., 1994, *Handbook of Usability Testing* (Wiley).

SANDERSON, P.M., and FISHER, C., 1994, Exploratory sequential data analysis: foundations. *Human-Computer Interaction,* **9(3)**, 251-317.

SAWHNEY, N., and SCHMANDT, C., 2000, Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments. *Transactions on Computer-Human Interaction*, **7(3)**, 353-383.

SCHMIDT-BELZ, B. and POSLAD, S., 2003, User Validation of a mobile Tourism Service. Proceedings of HCI in Mobile Guides (Udine, Italy: in conjunction with Mobile HCI 2003).

SCHMIDT-BELZ, B., LAAMANEN, H., POSLAD, S., and ZIPF, A., 2003, Location- based mobile tourist services - first user experiences, ENTER 2003 Conference, January 2003 (Helsinki).

SMITH, J., MACKANESS, W., KEALY, A., and WILLIAMSON, I.P., 2004, Spatial Data Infrastructure Requirements for Mobile Location Based Journey Planning. *Transactions in GIS*, **8(1)**.

STRAUSS, A., and CORBIN, J., 1998, *Grounded Theory in Practice* (Sage Publications).

UMLAUFT, M., POSPISCHIL, G., NIKLFELD, G., and MICHLMAYR, E., 2003, LoL@, a Mobile Tourist Guide for UMTS, *Information Technology & Tourism*, **5(3)**, 151-164.

VETERE, F., HOWARD, S., PEDELL, S., and BALBO, S., 2003, Walking Through Mobile Use: Novel Heuristics and Their Application. Proceedings of OzCHI 2003 (Brisbane, Australia: CHISIG).

WHARTON, C., RIEMAN, J., LEWIS, C., and POLSON, P., 1994, The Cognitive Walkthrough: A practitioner's guide. In *Usability Inspection Methods*, edited by J. Nielsen, and R. L. Mack (John Wiley and Sons, Inc,) pp. 105-140.

APPENDIX: MERGED PROBLEM LIST

Critical problems

1. **Maps.** Issues related to how the user interprets and uses maps in conjunction with the textual information.
2. **Navigation.** Issues related to problems with navigating through the screens of the system.
3. **Information.** Issues related to lack of relevance and accuracy of information presented by the system.
4. **System vs. World.** Issues caused by disparities in the relationship between information in the system and information in the world.
5. **Use and usefulness.** Issues related to a conception of use broader than usability (ISO) including the overall purpose of the device (e.g. social, lifestyle etc.)

Serious Problems

6. **Input and affordances.** Issues emerging from difficulties with entering data into the system and the affordances offered by the system for doing so.
7. **Help and recovery.** Issues related to the support lack of offered by the system and its inability to assist the user in recovering from errors.
8. **Knowledge about city.** Issues related to high requirements for user's knowledge about the city in which they are interacting with the system.
9. **Labeling.** Issues caused by poor wording and use of abbreviations within the system.
10. **Cognitive Load.** Issues related to high requirements for cognitive resources (memory and attention) to be able to use the system.
11. **System.** Issues caused by technical malfunctions in the prototype system.
12. **Interface flexibility.** Issues related to lack of support for variation from the predefined path of interaction.
13. **Mental model.** Issues related to disparities between how the system works and how the users think the system works
14. **User Confidence.** Issues related to lack of confidence in using the system or acting according to the information provided by the system.
15. **Scope.** Issues related to uncertainties regarding what functionalities the system offers to the user.
16. **Value.** Issues related to users experiencing limited value of the information presented by the system.

Cosmetic problems

17. **Efficiency.** Issues emerging from users experiencing the system being time consuming and cumbersome to use
18. **Orientation**. Issues emerging from lack of information in the system for supporting the user's orientation in the real world.
19. **Readability.** Issues related to difficulties with reading small fonts on the screen of the device.
20. **Dependency on the System.** Issued related to the user being dependant on the system for making decisions
21. **Social comfort.** Issues related to how comfortable the user is with using the system in public, with particular reference to the acceptability of using the system.
22. **Emotional response.** Issues causing strong emotional responses from the user while using the system.