Developing of a Qualitative Classification Method for Usability Errors after Rasmussen.

Adrian Haar – s1002457

20.02.2013

Supervisors: Dr. M. Schmettow, Prof. Dr. J.M.C. Schraagen

UNIVERSITY OF TWENTE.

Abstract

The study reported in this paper investigates the problems related to the classification of usability problems into the three categories, skill- rule- and knowledge-based errors, introduced through Rasmussen (1983). Classifications of usability problems were conducted in prior studies but most of them lack a specific description how this classification was executed. We suggest that a reliable method is required to classify usability problems. For this purpose we developed a decision schema, which we evaluated through analysis of inter-rater agreement of three raters classifying usability problems, gathered in an experimental usability testing study. This analysis brought no sufficient evidence for the reliability of this schema. As no systematic flaws in the developed schema could be identified, we suggested that the raters did not possess sufficient information to conduct a reliable classification. We conducted a qualitative analysis of all incidents provided through the usability evaluation to investigate which information is needed by raters to classify usability problems and how the method of data gathering can be adapted to provide this information. The findings of this analysis support the assumption that the raters did not posses enough information and further showed that the intention of the user is one of the most important pieces of information that needs to be gathered through the evaluation method and reported to the raters to ensure a reliable categorization of usability problems. Further does the results of this study stress the need of a reliable method to classify usability problems.

Abstract

De studie beschreven in dit artikel onderzoekt de problemen in verband met de classificatie van usability problemen in de drie categorieën, skill- rule- en op knowledge- based fouten, geïntroduceerd door Rasmussen (1983). Classificaties van usability problemen zijn in veel eerdere studies uitgevoerd, maar een specifieke beschrijving, hoe deze classificaties zijn uitgevoerd, wordt bijna nooit gerapporteerd. Wij stellen dat een betrouwbare methode nodig is om usability problemen te classificeren. Hiervoor ontwikkelden we een beslissingsschema, dat we evalueerden door de inter-beoordelaar overeenkomst van drie beoordelaars te analyseren, die in een experimentele usability testing verzamelde usability problemen classificeerden. Deze analyse bracht geen voldoende bewijs voor de betrouwbaarheid van dit schema. Aangezien er geen systematische fouten in het ontwikkelde schema geïdentificeerd konden worden, suggereren wij dat de beoordelaars niet over voldoende informatie beschikten om een betrouwbare classificatie uit te voeren. We voerden een kwalitatieve analyse van alle incidenten uit die via de usability evaluatie gevonden worden om te onderzoeken welke informatie voor beoordelaars nodig is om usability problemen te classificeren en hoe de methode van dataverzameling kan worden aangepast om deze informatie te verkrijgen. De bevindingen van deze analyse ondersteunen de veronderstelling dat de beoordelaars niet over genoeg informatie beschikten en tonen verder aan dat de intentie van de gebruiker een van de belangrijkste informatiestukken is dat via de evaluatiemethode moet worden verzameld en verder aan de beoordelaars gerapporteerd moet worden om ervoor te zorgen dat een betrouwbare classificatie van usability problemen mogelijk is. Verder onderstrepen de resultaten van dit onderzoek de noodzaak van een betrouwbare methode om usability problemen te classificeren.

Table of Contents

Introduction

The development of new technologies was never this fast and the target group of these technologies was never so broad. Because of the variety of users, usability testing became necessary to guarantee that the masses of users can work with a given product. Compared to other disciplines usability testing and evaluation is a young discipline and its methods are not yet fully developed. Testing the usability of a new product produces lots of data that have to be analyzed. Quantitative usability testing methods, such as time-on-task analyses, provide established routines to cope with this flood of information. On the other hand, in qualitative methods, like think-aloud transcripts and video recordings, the methods used to analyze the given data are often inconsistent (Gorlenko & Englefied, 2006).

Current and common practice in usability testing is to observe novice and/or expert users to find problems with the usability of a product. An advantage in comparing users at different skill levels is that they not only differ in the number of errors they produce, but also in the qualitative nature of the problems they experience. Novice users usually encounter significantly more and more critical problems than experts, while experts experience more cosmetic problems (Kjeldskov, Skov, & Stage, 2005). Due to this finding a combination of novice and expert users should be used as subject population / target audience for a consistent usability test.

For new products, yet to be released to the market, this recommendation is obviously not feasible, as there are no expert users available for testing yet. In this case usability testing is used to investigate if novice users can interact intuitively with the new product or if the product has to be changed to fulfill this purpose. However,

this tactic / procedure ignores the fact that users will learn from the mistakes they make as they interact with the product and will probably change their behavior over time. Hence, it seems that current usability studies tend to evaluate the discoverability of a certain feature rather than the usability concern per se, which would remain problematic in continued use of the product (Courage, Jain, & Rosenbaum, 2009).

This problem could be avoided through the availability of more qualitative information about the nature of usability problems and their development over time. Users gain more experience with a given product and therefore learn how to use this product in the way the developer intended and to cope with given situations. This phenomenon does not get a lot of attention in research because it can only be investigated further through longitudinal research designs. Until now, longitudinal research designs in usability research are rare but would provide valuable information. As example could information be gathered about how measures of effectiveness and satisfaction of the user, with a given product, would develop over time (Hornbæk, 2006).

Usability problems are generally classified as being skill-based, rule-based or knowledge-based (Fu, Salvendy, & Turley, 2002). This classification, based on the SRK- framework developed by Rasmussen (1983), refers to the degree of conscious control exercised by an individual over his or her activities that produces the error/problem. An activity that is indicated to be knowledge-based will be executed with high conscious monitoring needed. Actions on the knowledge-based level mostly occur with a novice performing a task or with an experienced individual in a complete novel situation. A skill-based task and represents a highly practiced task that can be executed with nearly no conscious monitoring needed (Embrey & Lane, 2005).

Such a categorization of usability problems has a lot of advantages in the qualitative analysis of a given product. Past research has shown that as users gain more experience in the use of the product, some errors, that the user learned to avoid, disappear, but also new errors emerge (Kjeldskov et al., 2005). It can be hypothesized that this phenomenon lays in relationship with the category the given error is associated with and that knowledge-based errors disappear faster than skill based errors. This hypothesis has to be tested in further research. If this hypothesis holds, a reliable categorization of usability problems could provide a more effective way of product evaluation. For instance could be prevented that effort will be invested in correcting usability problems, which might disappear over time.

Earlier research found that, in usability evaluations, experts will detect more usability problems that are associated with skill- and rule-based errors and that novice users will find more problems at the knowledge-based level (Fu et al., 2002). This finding suggests that novice users learn how to avoid knowledge-based errors, but that knowledge- and rule-based problems remain, or even emerge, as they gain more expertise.

As mentioned earlier, qualitative usability research struggles with the consistency in the used methods (Gorlenko & Englefied, 2006). In addition to this, studies of usability evaluation tend to focus on measures as the numbers of problems identified and rarely describe the process of evaluation (Norgaard & Hornbæk, 2006). Next to the analysis of think aloud transcripts and video recordings, also the categorization of usability problems is an example for this problem. The research of Fu and colleagues (2002) did already use the categorization of Rasmussen (1983) in their experiment and analyses but did not report how they categorized the problems. This could be a threat for the replicability of their findings: If the categorization of the

problems is not replicable, further findings about the categories themselves could be unreliable.

Developing and testing a taxonomy to classify errors in human-computer action Zapf and colleagues (1992) performed a field experiment. In this field experiment observers watched office workers performing their tasks and reporting the observed errors. Relying on these reports, the different errors were differentiated in classes. Similar to the article of Fu and colleagues (2002) Zapf and colleagues (1992) do not report how this classification is conducted.

Likewise, the European Committee for Electrotechnical Standardization uses the taxonomy of Rasmussen (1983) in their international standard of application of usability engineering to medical devices, but does not provide guidelines on how to classify a given usability problem as being skill-, rule-, knowledge-based (European Committee for Electrotechnical Standardization, 2008). This leak on information could also lead into confusion about the nature of a given problem.

This illustrates the need for more standardized methods to categorize usability problems. In qualitative usability research, standardized protocols for the analysis of think-aloud protocols or video recordings are available. For the categorization of usability problems, however, no such method is available, yet.

The purpose of the study reported in this paper is to develop a method to reliably categorize usability problems into the three error categories developed by Rasmussen (1983). This method was mostly deduced from publications of Reason (1990) who worked further on the framework established by Rasmussen. To test this method, different raters categorized usability problems according to the SKR-framework using the developed method.

Method

General Study Design

We developed a method to classify usability problems based on the classic works of Rasmussen (1979) and Reason (1990). This method was tested with real usability problems, found in different usability tests. We then had naïve subjects categorize these problems using aforementioned decision method. By calculating inter-rater agreement, the reliability of this method was probed, lending high reliability measures if the raters categorized the problems in the same way. Based on the results of this analysis the method was adjusted.

In the second experiment, new problems were gathered with a broader choice of applications to test if the applicability of the adjusted algorithm would also extend to applications outside the personal management domain. Inter-rater agreement was again calculated to analyze the reliability of the categorizations.

Development of the method

In order to develop a method to categorize usability problems, we used theories and models grounded in the field of human error. An early method to classify human errors was to distinguish between slips and mistakes. Errors are defined as slips if the plan of actions to solve a task was correct but the execution was flawed. By contrast, errors are defined as mistakes when the intended action was not appropriate but the execution was right (Norman, 1981).

Reason (1990) published an algorithm (see Figure 1) to determine if a certain error was a slip or a mistake, which will be used as basis for the classification algorithm for usability problems.

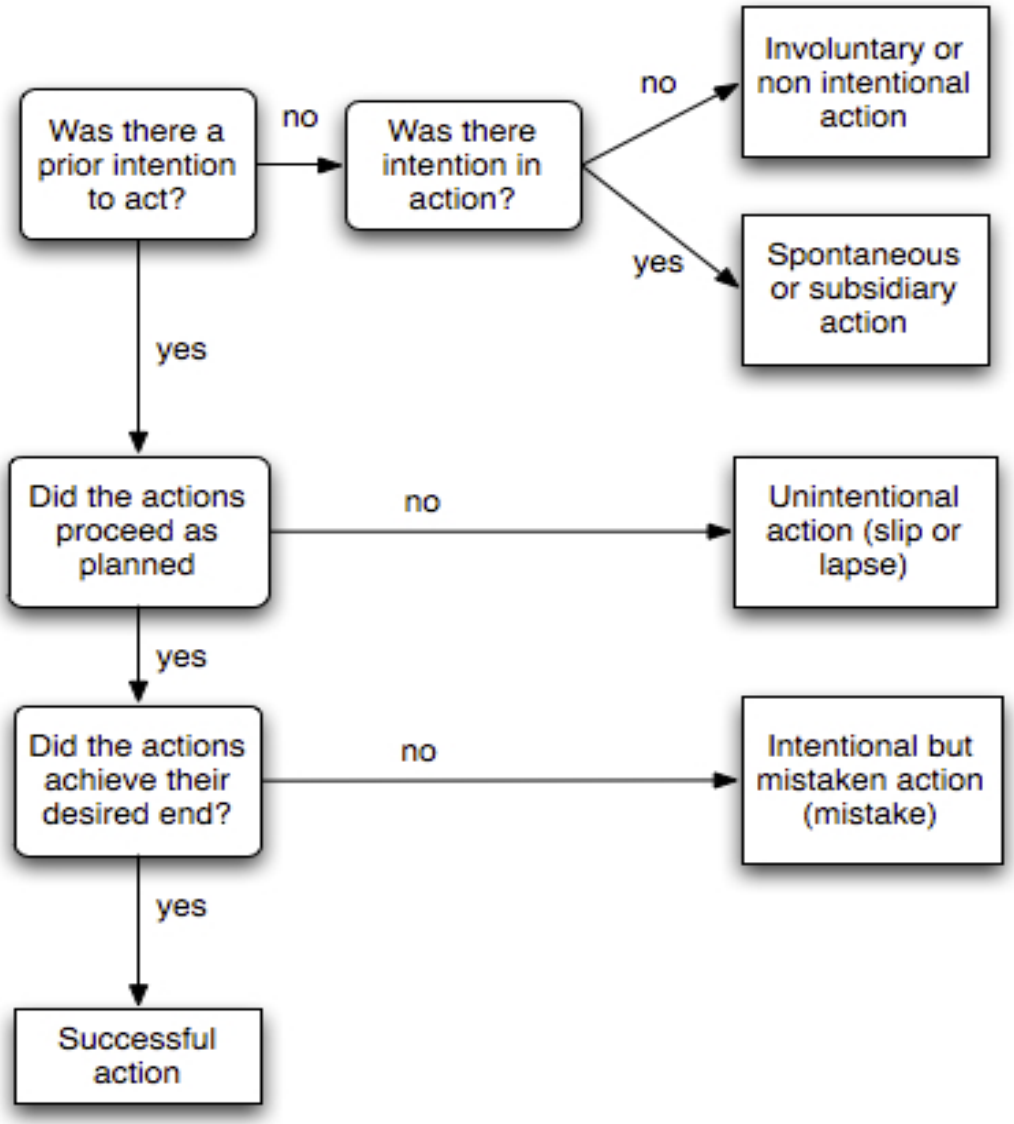


Figure 1. Reason's algorithm for distinguishing the varieties of intentional behavior. The three main categories are non-intentional behavior, unintentional behavior (slips and lapses) and intentional but mistaken behavior (Reason, 1990)

According to Reason (1990) the error types, slips and mistakes can be related to the SRK- framework of Rasmussen (1983). Table 1 illustrates that slips can be related to errors on the skill-based level and mistakes with errors on the rule- and knowledge-based level.

Table 1. Relating the three basic error types to Rasmussen's three performance levels (Reason, 1990)

Performance level	Error type
Skill-based level	Slips and lapses
Rule-based level	RB mistakes
Knowledge-based level	KB mistakes

Reason adds another type of error he calls lapses. In this paper we will discuss the details of this distinction, since both types can be related to the skill-based level of the SKR- framework.

With this additional information, the algorithm can be adapted to distinguish between skill-based errors, knowledge- and accordingly rule-based errors. On the one hand, if we can classify an error as a slip we can conclude that this error could also be classified as an error on the skill-based level. On the other hand, errors classified as mistakes, in accordance with the algorithm after Reason (1990), could also be classified as rule- or knowledge-based. These adaptations of the original algorithm can be found in Figure 2.

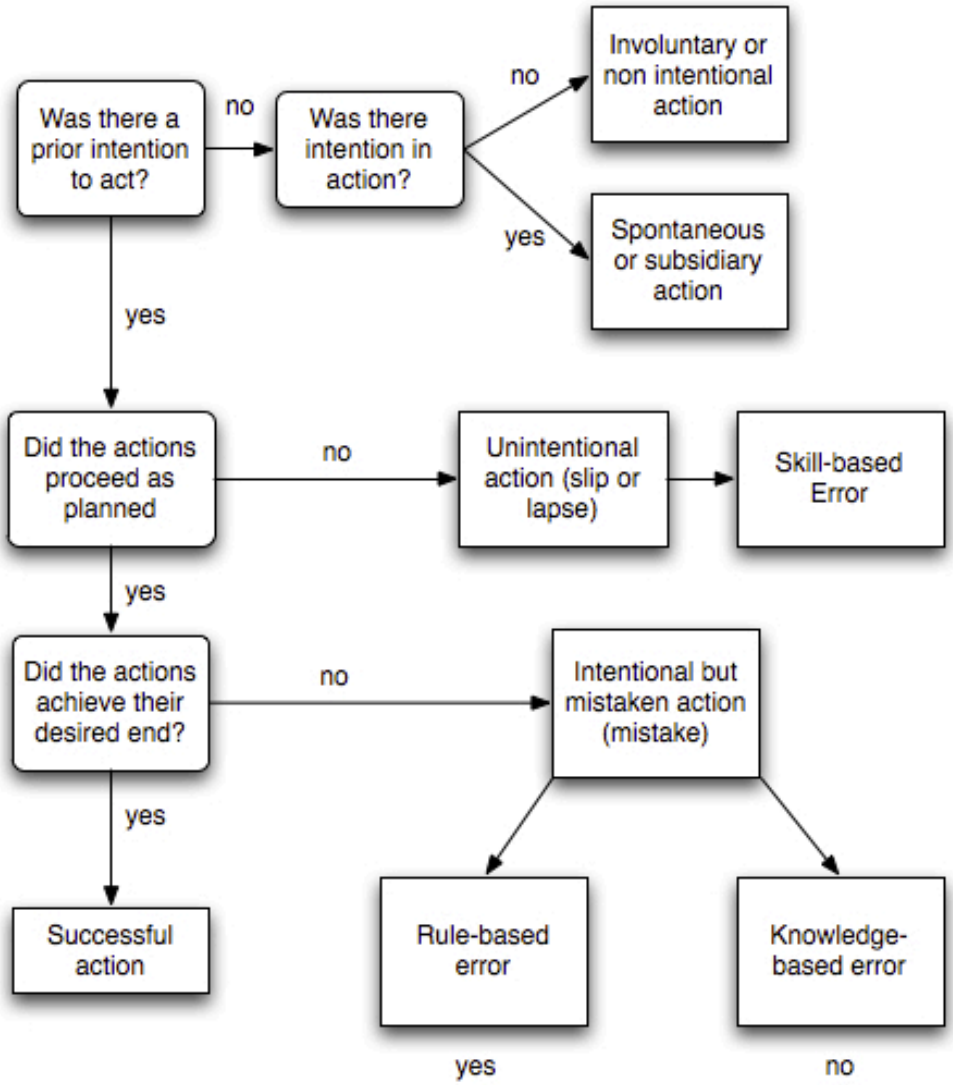


Figure 2. Adapted Algorithm with the SKR- framework added

This adapted algorithm lacks a distinction between knowledge- and rule-based errors. In the definition of the SRK- framework, Rasmussen (1983) defined that rule-based errors are typically related to the use of an if-then statement. A rule-based error typically entails a misinterpretation of the situation or incorrect recall of procedures. An error on this level would imply that the individual applied the wrong rule for the situation or the individual misinterpreted the situation and would therefore apply an inappropriate rule. On the knowledge-level individuals do not work with such if/then rules, because for example they did not solve a comparable situation earlier to

establish an applicable rule. With this additional information about the distinction between rule- and knowledge-based errors we can adapt the algorithm. This algorithm, shown in Figure 3, can be used to work with a problem as input and gives a classification in accordance with the SKR- framework of Rasmussen as output.

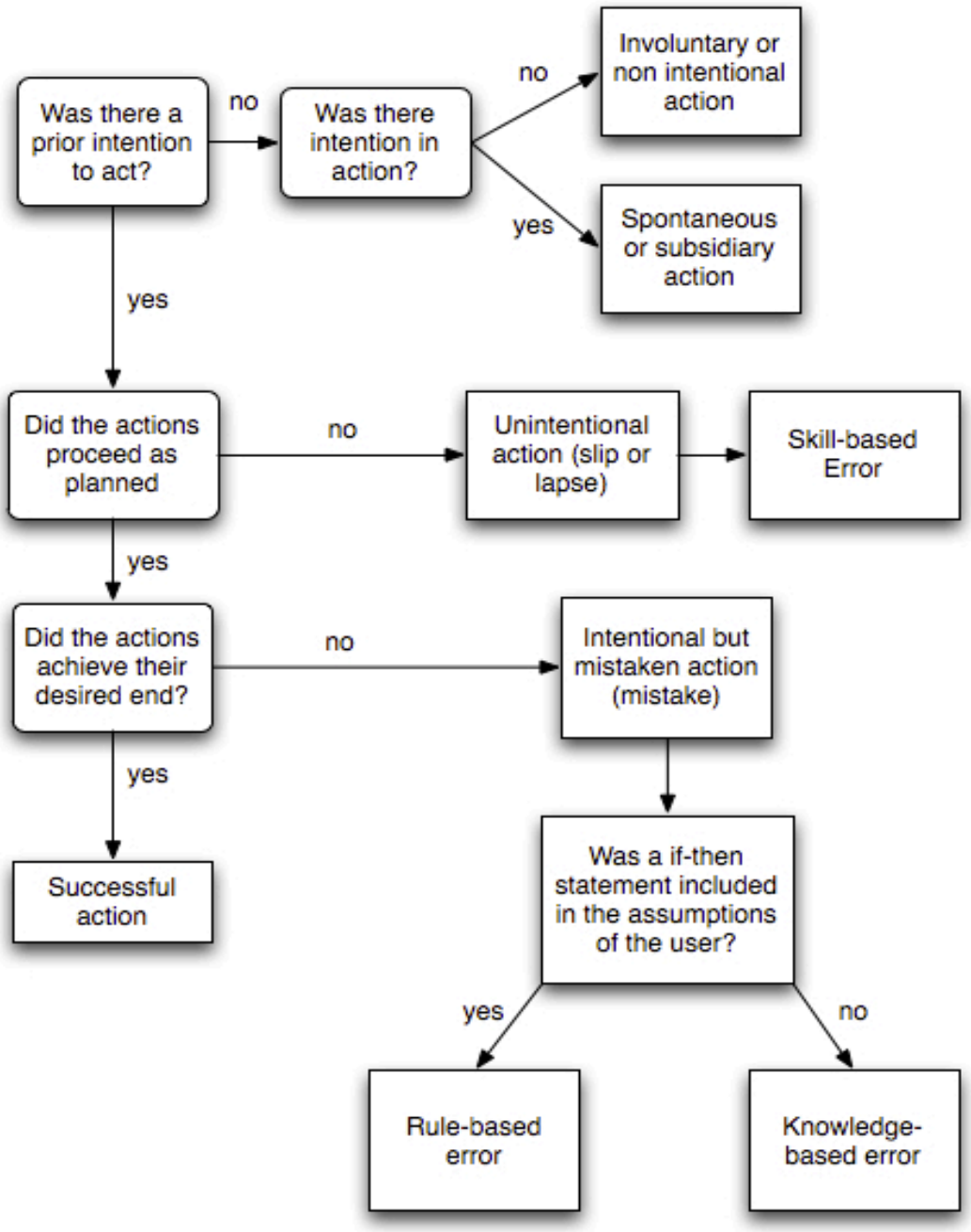


Figure 3. Algorithm with added distinction between rule- and knowledge-based errors

Experiment 1

Procedure

In a first experiment 24 psychology master's students following the "Human Factors and Media" track were subdivided into five groups. Each group had the assignment to test a different personal management application. Also, every student had to participate in two tests of another group.

The students used applications developed for Android, which were emulated on a standard home computer. On this computer the usability testing software Morae was used. To achieve consistency in the testing of the different groups, each group was provided with the same tasks and identical think-aloud, matching and problem reporting instructions. The matching and problem reporting instruction were adjusted from the method for structuring usability problem reports by Lavery, Cockton, and Atkinson (1997), as reported and evaluated by Hornbæk and Frøkjær (2008) in comparison to other matching methods.

Subsequent to the testing, the problems found in the usability evaluation have been collected in a catalogue. This problem catalogue, consisting of 12 problems and the developed algorithm, were handed to three independent raters who categorized the problems with the help of the developed algorithm. These raters were master students or graduates from different fields to ensure that not only experts in the field of usability would work with this algorithm. This ensures that the raters really use the algorithm and not relying on their experience or other methods they were using (Besnard & Cacitti, 2005). The different raters were only instructed to read the reports about the different usability problems and use the given algorithm to categorize the problems into the three different categories of the SKF- framework.

Results

The classification of the given usability problems according to the different raters could be found in Table 2. The different color shading illustrates that agreement in only five of the twelve problems can be found.

Table 2. Classification of the given usability problems into the categories skill-based (SB), rule-based (RB) and knowledge-base (KB)

Problem ID	Rater A	Rater B	Rater C
#1	RB	RB	RB
#2	RB	RB	RB
#3	KB	KB	RB
#4	SB	KB	SB
#5	RB	RB	RB
#6	KB	SB	RB
#7	KB	RB	RB
#8	KB	KB	KB
#9	SB	SB	KB
#10	SB	SB	SB
#11	SB	SB	RB
#12	SB	SB	KB

With the purpose to get more insight into the agreement of the three raters and the inter-rater reliability, we conducted different statistical methods. We calculated a Cohen's Kappa value for each pair of raters, to determine if relation between the single raters could be found. Also, the intraclass correlation coefficient (ICC) was calculated to get more information about the over-all agreement. The ICC, is next to the more popular Fleiss Kappa, a measure of the reliability of ratings and can be used to access the inter-rater reliability of more than two raters (Shrout & Fleiss, 1979).

Between Rater A and Rater B a Kappa value of .621 with a significance of .002 could be measured. Rater A and Rater C had a Kappa value of .287 with a significance of .097 and Rater B and Rater C a Kappa value of .258 with a significance of .139. The ICC calculated for the sample of raters resulted in a value of .413 with a significance of .139.

Table 3. Results of the statistical agreement analysis

Statistical Method	Measure of Agreement	Significance
Kappa Rater A * Rater B	0.621	0.002
Kappa Rater A * Rater C	0.287	0.097
Kappa Rater B * Rater C	0.258	0.016
Intraclass Correlation Coefficient	0.413	0.139

Conclusion

The inter-rater correlations do not show a sufficient measure of agreement and also only one significant result. However, we cannot isolate a specific fault in the algorithm through this data. If all raters would agree about the classification of skill-based errors and we only had noise in classification of rule- and knowledge based errors we could assume that this distinction in the algorithm is not sufficient.

Therefore, that we cannot dietetic a defect in the algorithm, we try to increase the measure of agreement between the raters in the following experiment. To achieve this goal the raters will receive an instruction text with step-by-step explanation to each point of decision in the algorithm. Further, the function of the algorithm will be

illustrated with an example out of our pre-test. Through this addition of training we assume better values of agreements in the next experiment.

Experiment 2

The following experiment is designed to gather usability problems of mobile applications to test the applicability of the developed algorithm to broader range of applications. A different set of participants was asked to solve a pre-specified set of tasks in a selection of different applications. Afterwards, they were interviewed afterwards to recall their thoughts during usability incidents they encountered during the experiment (retrospective think-aloud). The participants were all students of the University of Twente and were not selected for any given criteria. Four participants of this test were female and six were male. Age was ranged between 20 and 27 with a mean age of 23.3.

Procedure

The participants worked with four different applications developed for Android smartphones and tablets. All of these applications served different purposes, for instance finding a restaurant or checking the weather. The choice of applications had the intention to assure that the developed algorithm is applicable to other applications than the personal management apps used in the pre-test. All applications were emulated with BlueStacks on a laptop, to provide the possibility to record the screen and the face of the participant at the same time. This laptop ran the usability testing software Morae with the possibility to record the screen of the laptop, a video

recording of the user and audio. Further, Morae was used to connect an observer laptop. On this observer laptop, the instructor of the experiment made notes and observed the participant working on the given tasks. He also marked usability incidents, which should later be matched into usability problems, during the recording through an observer. These notes were send to the participant's laptop and saved together in one file on the participant's laptop. A schematic overview of the used experimental setup can be found in figure 4.

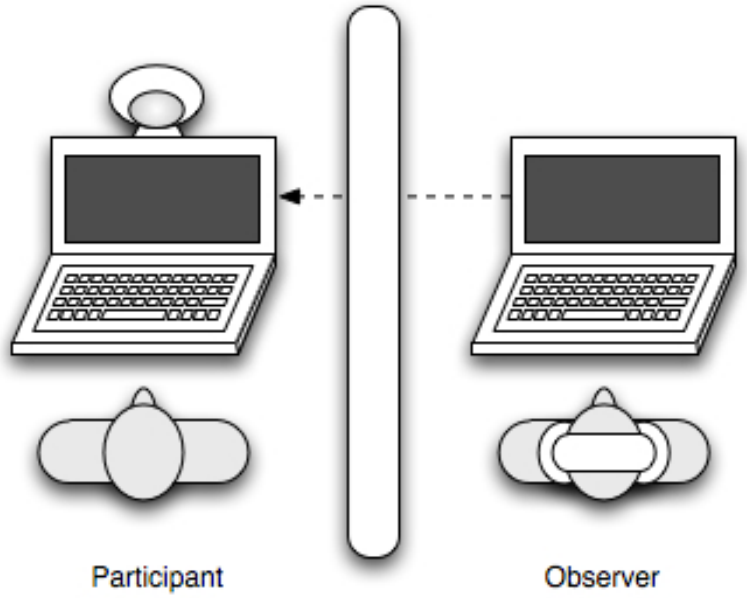


Figure 4 - Schematic overview of the experimental setup

The participants completed a list of tasks handed to them on a sheet of paper. After completion, the screen capture and audio/video log of their performance was played back to them. While watching, they were ask to repeat out loud the thoughts they had during execution of the task and especially when they encountered problems.

This technique, retrospective think-aloud, reduces the participants' cognitive load during the task, therefore we could assume that the errors occurring during the test are not related to the extra task to verbalize his or her thoughts (Hertzum, Hansen,

& Andersen, 2009). Further it is possible to prompt the participant during retrospective think-aloud for more qualitative information about the problems they experienced as with classic think-aloud technics. The combination of a recording of the actual interaction of the participant with the tested product and a retrospective interview fits the suggestion Nielsen and Yssing (2004) formulated for the procedure of usability testing.

A total of 111 usability incidents were logged, listed, and matched to usability problems according to the method of Lavery et al. (1997). The resulting problem catalogue of 15 usability problems was again given and explained to three independent raters, who rated each problem using the supplied algorithm and the afore-described additional information and training material. Agreement will again be assessed by calculating kappa and ICC indices, to evaluate the reliability of the developed algorithm.

Results

Subsequent to the classification each rater reported their classification. Rater A classified six problems as skill-based, one as rule-based and eight as knowledge-based. Rater B assigned two problems as skill-based, four as rule-based and nine as knowledge-based. Rater C categorized eight problems as skill-based, six as rule-based and one as knowledge-based. Agreement between all three raters was reached in only two of all 15 problems (Table 4).

Table 4. Classification of the given usability problems into the categories skill-based (SB), rule-based (RB) and knowledge-based (KB)

Problem ID	Rater A	Rater B	Rater C
#1	SB	KB	SB
#2	SB	KB	SB
#3	KB	KB	RB
#4	KB	SB	SB
#5	KB	RB	RB
#6	SB	KB	SB
#7	KB	RB	SB
#8	KB	KB	SB
#9	KB	KB	RB
#10	KB	KB	RB
#11	SB	KB	SB
#12	SB	SB	SB
#13	RB	RB	RB
#14	SB	RB	RB
#15	KB	KB	KB

To get a better insight about the agreement of the three raters, different statistical methods were applied. As in the first experiment we calculated the Kappa value for each combination of raters in pairs and the intra-class correlation coefficient (ICC) for all three raters.

Between Rater A and Rater B a Kappa value of .124 with a significance of .445 could be measured. Rater A and Rater C had a Kappa value of .264 with a significance of .032 and Rater B and Rater C a Kappa value of .233 with a significance of .042. The ICC calculated for the sample of raters resulted in a value of .419 with a significance of .108.

Table 5. Results of the statistical agreement analysis

Statistical Method	Measure of Agreement	Significance
Kappa Rater A * Rater B	0.124	0.445
Kappa Rater A * Rater C	0.264	0.032
Kappa Rater B * Rater C	0.233	0.042
Intraclass Correlation Coefficient	0.419	0.108

Conclusion

The results show that the tested algorithm does not provide a reliable basis for categorizing usability problems in terms of the SRK-framework. Normally, a Kappa or ICC value above .6 could be interpreted as a reasonable degree of agreement but cannot be found in the results of the current study.

The first interpretation of these results could be that the algorithm has flaws and does not suit the purpose to categorize usability problems. Unfortunately, the data does not allow for conclusions about the nature of these flaws and how the algorithm could be adjusted.

If the results would show that all raters could clearly distinguish slips from mistakes and therefore skill-based errors from rule- and knowledge-based errors, we could conclude that the added distinction is not explicit enough. All other parts of the algorithm rely directly on the publication of Rasmussen (1983) or Reason (1991).

Through the additional training provided in the second experiment we also can assume that the raters did understand the given method and did not have problems with the application of the method itself.

Given that the raters need a detailed idea of the situation of each usability problem, another source of error could be the way in which usability problems are presented to them. In our experiments we used a method developed by Lavery (1997) and adjusted by Hornbæk (2008). Hornbæk adjusted Lavery's method to provide a manner of reporting usabilty problems that does not require complex knowledge about the theories behind usability testing. The intention of these reporting methods was to give the developers of the evaluated product a concrete idea of the problem, so that they could change the product in a way to solve the discovered problem.

In the current experiments, however, the focus lays on the user part instead of the product and the reports were intended to capture the thoughts of the participant and why he conducted the actions the way he did. It is quite possible that the raters therefore did not dispose of enough information of this kind to reliably classify the given usability problems.

Hence the question arises: What information do raters need to reliably categorize a usability problem within the SRK-framework?

Qualitative Analysis

To get a more detailed view on the information needed by the raters we decided to conduct a deeper, quantitative analysis of the raw data gathered in the second experiment.

Usability evaluations usually produce much more usability incidents than can be handled by the designers individually. Matching is therefore used to condense this large amount of usability incidents to a more manageable amount of usability problems that can subsequently be tackled jointly. Following a similar rationale we used matching to bring back the large amount of 111 observed usability incidents to a reasonable amount of incidents pointing to common underlying problems. This yields the advantage of an amount of information manageable by our raters.

On the downside, through the process of matching and reporting of the resulting usability problems, some information is, naturally, filtered out and, hence, lost. The sort and/or amount of information that is lost does not seem to pose a problem for usual usability evaluations. However, the lack of inter-rater agreement in the current study points to the possibility that the sort of data that gets filtered out is crucial for the raters' ability to reliably classify the problems into the SKR-framework. Additionally, information might get lost through the used method of reporting: as discussed earlier Hornbæk's technique focusses on the technical rather than the human side of usability problems, which is more important in the current situation. Alternatively, it is possible that insufficient information of this kind was gathered through the retrospective think-aloud interview in the first place.

To investigate if crucial information was indeed not collected or whether it was lost at the later matching and reporting stage, we examined the totality of raw data. To this end, we first overlaid the video recordings and screen captures of the

participants performing the task with the recording of the retrospective think-aloud interview, lending a simultaneous picture-in-picture recording. Thereby, we arrived at a single recording combining both the data of the execution stage, provided by the screen captures and the facial expressions of the participant, and data on the thoughts and intentions of the user, as reflected in the RTA data.

Based on these recordings, we then examined whether the combined information was sufficient to classify each individual usability incident using the information provided by the newly developed classification schema and the information provided in the original publication by Rasmussen (1983). If, following this method, a classification is not possible the crucial information apparently was not collected through the gathering method used in the experiment. On the contrary, if the classification is possible, the crucial information is provided through the gathered data, but must have been lost at a later stage, likely during matching and reporting to the raters.

Skill-based Incidents

In general, Rasmussen (1979) defines a skill-based action as an action that is not in accordance with the intention of the individual. A first incident that illustrates this definition could be observed during the experiment. The participant tries to scroll a list of results in an application that searches for restaurants. He accidentally selects a restaurant and comments in the retrospective think-aloud interview: *"Okay, here I'm selecting a restaurant but did not actually want to"*[1]. If we only had the video recording of this incident it would have been hard to conclude if the action was

[1] This and all following quotes were translated from German to English because the participants were allowed to use their mother language in the retrospective think-aloud.

intended or not. Seen in combination with the information provided by the retrospective think-aloud we knew that he intended to scroll the list, rather than to select a restaurant. With the combination of observational and think-aloud information this incident could therefore confidently be classified as skill-based.

Another incident related to scrolling happened, as a participant wanted to get more information about a restaurant and tried to scroll its profile site. He accidentally hits the "Call the restaurant" button in place of scrolling down. The participant does not comment this incident, but a second participant, experiencing the same problem, says: *"Oh, here I clicked wrong"*. To classify these incidents the information given through the retrospective interview is not as important as in the first incident, because it is obvious that the participants did not want to call the restaurant in the experiment: It was not part of the given tasks and the participants should know that this function was not given in the experimental setup. Hence we can safely assume that this incident occurred against the intention of the participants and therefore can be classified as skill-based errors.

More incidents related to scrolling and accidental clicking can be observed but all of them are comparable to the two illustrated above. Either the participants verbalize that this action was in conflict with their intention (e.g., *"That was not on purpose", "Oops, I clicked wrong"*) or the resulting action was in conflict with the given task.

Another set of incidents that was frequently observed relates to the "Back" and/or "Menu" button. For example, one participant was finished with a task and wanted to go back to the main menu of the application but hits the "Back" button too often and therefore accidentally closes the application. Likewise incidents can be observed with nearly all participants, but only one of them commented on this

incident during the retrospective interview (*"I don't know why I did this"*). Although, in this case, the RTA alone does not point to a usability problem in the remaining users, the screen recordings showed that all participants quickly re-started the application following likewise incidents. It therefore seems safe to assume that, similar to the subject who actually commented on this incident in the RTA, they had no intention to close the application. These incidents can therefore be classified as skill-based.

This rationale did not hold for a single incident, however, in which the user closed the application, waited a moment and re-started the application, but it can also observed through the facial camera that the participant reads the instructions again and also states in the interview: *"I thought I was finished with the application but then it says in the instructions that the next task had be solved with this application again."* With this additional information we can conclude that in this case the closing of the application was intentional and therefore not a skill-based error. The important difference in this incident is that the participant does not behave against his or her intention. This conclusion only can be stated through the combination of all additional information gathered through the retrospective interview and the video recording.

When using the weather application a participant can be observed committing another error that can be classified as skill-based. She tries to search for her hometown to gather the weather information for the next day, but as she is finished with typing the name, she does not click the search button so the results are not updated. The user comments this in the retrospective think-aloud with: *"Oh I forgot something here."* Without this additional information it would be hard to classify this incident because she also could have misunderstood the search function to work without hitting the search button but through the fact that she only forgets to hit the

search button it could be concluded that she was acting against her intention and therefore this incident can also be classified as skill-based.

For the classification of the next incident the additional information of the interview was also needed to arrive at a reliable classification. In a Twitter client, the user tries to search for a band he wants to follow but clicks another option. With this observation alone it would be hard to classify this incident because it cannot be known if this was the intention of the user because he did not know where to search for the band. In the retrospective interview the user says, *"Oops, there I clicked the wrong option accidentally"*. With this additional information we know that the user did know which option to choose to search for a band and also had the intention to click this right option but accidentally clicks the wrong one. Therefore this incident can also be classified as skill-based.

Rule-based Incidents

Rule-based errors are categorized by Rasmussen (1979) as errors that are related to the wrong application of a certain rule or assumption. In contrast to skill-based errors, individuals execute their intention right but the intention was not chosen right to solve the problem. To categorize an incident as rule-based the available information should indicate that the participant thought that if he/she would commit a certain action he/she would get a certain result but that the actions based on this assumption did not lead to a correct solution of the problem.

One rule-based incident could be observed with four participants working with the weather application. The participants were trying to change the location for the weather by dragging a pointer on a map to his hometown. This feature is not supported by the weather application but is common practice in other web services for

example Google Maps. One participant comments this incident in the retrospective interview with: *"I'm trying to change the location by dragging this red thing but it does not work."* Taking this information you could formulate the intention of the participant as: "If there's a map with a pointer I can drag this pointer to change the location." This illustrates the application of a working rule but in the wrong context. Therefore we can classify this incident as rule-based.

Another wrong application of a working rule could be observed among a participant working with a twitter application. The task was to change the name of the given account. One participant has problems with finding the right menu to change the name and comments this in the retrospective interview: *"I have some problems here because I thought that I have to change the name in the settings menu."* The participant makes the assumption "If I want to change the account-name this option will be offered in the "Settings" menu." This rule will work in many other applications but in the tested twitter application the name can be changed in the "Profile" menu. Therefore the participant applied a working rule in the wrong context and therefore this incident could be categorized as rule-based.

A wrong application of working rule can also be observed in the following incident. The participant was working with the Twitter application and was finished composing a tweet, but has problems sending it. The participant comments this in the retrospective interview: *"I was trying to send the tweet via the enter button because I'm used to this (...)".* The rule that the participant learned from other applications and tries to adopt here, could also be reformulated as: "If I want to send a message, I can use the enter key." This assumption would potentially work in other applications but not in the one tested, so we classified this incident as rule-based.

The next incident was also observed with the tested weather application and the participant also applied information she learned from her past use of other applications. The participant was trying to find more information in the application but was only looking for it on the main page. She commented this in the retrospective interview with: "*I was thinking that you cannot click the buttons* [representing the other pages of the application] *because they did not have the same shading as the other buttons; if buttons are grey you cannot click them"*. The participant thereby formulated the wrongly applied if/then rule himself. The incident can therefore be classified as rule-based.

Working with the restaurant-finder application a participant could be observed to experience another incident that could be classified as rule-based. The participant was clicking the picture of the selected restaurant and displayed a confused facial expression. She comments this in the interview as following: *"There I thought that I would get more information but got only this* [the picture of the restaurant]*."* With this information we can formulate the intention of the participant as "If I click on the picture, then I will receive more information." Assuming that this expectation stems from the participants past use of other applications, we can conclude that this is another example for the misapplication of a working rule.

Knowledge-based Incidents

Knowledge-based behavior is mostly observed if an individual has to improvise in unfamiliar situations and if no rules or routines are available for handling a situation (Reason, 1990). Individuals do have a goal formulated through the task but lack a ready-made plan they can follow to achieve this goal. Therefore the plan how to solve the problems or a mental model for the situation has to be developed and

tested through trial and error (Rasmussen, 1983). The transition between rule- and knowledge-based behavior can be vague, as participants sometimes act after a plan of action in terms of if/then assumptions, but shift into knowledge-based behavior as this plan does not work out.

A first incident that can be classified as knowledge-based can be observed among two participants working with the radio application. The task formulates their goal to search for a radio station they usually listen to at home. The participants did not know how to solve this and therefore develop a plan for the situation, which does not really work out. One of the participants comments his or her actions in the retrospective interview as follows: *"I'm searching for a search field but I cannot find one, after some time I'm finding the magnifying glass in the upper right corner"*. The participant has developed the plan to search for a radio station using a search function and now tries to execute this plan. This plan fails at first due the fact that the participant cannot find a search function in the application. This trial and error behavior indicates that the observed incident can be classified as knowledge-based.

Working with the Twitter application, two participants could be observed behaving in a likewise trial and error manner. The participants' task was to compose a message but they experienced problems finding an option to get a keyboard. One of the participant comments this situation as following: *"Through clicking randomly I conjured a keyboard."* The other participant comments that he also *"clicked randomly"* without a real plan of action. The participants therefore had no plan on how to start typing and therefore discovered a working plan through trial and error, which classifies this incident as an incident the knowledge-based level.

The participant in the prior incident verbalized in the retrospective interview that he found a solution to achieve his goal through chance. This phenomenon could

also be observed in the following incident: Again, the participant was working with the Twitter application and was trying to complete the task to change the name of the twitter account. He did not really have a plan where to change the name and therefore searches the application through selecting different menus. The participant completes the task after a short time of searching and comments this in the retrospective interview: *"Now I complete the task but I have no idea how I did this."* This comment illustrates that the participant did not act according to a pre-defined plan of action but achieved the completion of the task only through trial and error. With this information we can classify this incident as behavior on the knowledge-based level.

Another incident that could be observed by eight of the ten participants was related to a task that should be solved using the weather application. The task was to change the unit of the temperature-scale from degree Fahrenheit to degree Celsius. All participants started to search the application's settings-menu for an according option. While most of the participants did get to the correct settings option, but did not recognize it as such as it was labeled with the words "Metric" and "Imperial", which is more commonly associated with the measurement of distance, not temperature. A participant comments this in the retrospective interview: *"There I found the right option but I did not know the meaning of this words"*. The participants could not interpret that this option would change the unit of the temperature from Fahrenheit to Celsius and therefore searched on for other options to achieve their goal. One participant comments this situation as following: *"At this point I'm clicking helplessly through the application to find an option"*. Four of the participants later came back to the right option to try what the function of this option is and solved the task that way. The other four participants gave up and did not complete this task. In the beginning the participants proceeded according to a plan on how to change the temperature but

as that plan did not work out all the participants shifted into a trial and error behavior to develop a new plan to achieve their goal. Therefore this incident can be classified as knowledge-based.

Working with the restaurant-finder application a participant could be observed experiencing another problem and behaving in trial and error manner. The task was to find reviews about a restaurant searched earlier. The participant tried to find the reviews and clicked some options available on the restaurant page. He comments in the interview that he *"misclicked"* on the different buttons. This statement would indicate that he did not click on the different buttons on purpose, which stands in conflict with the search behavior he displays. Therefore it is more likely that he tries to find the right option by trial and error and therefore we would classify this incident as knowledge-based and not as skill-based.

Similar to the prior incident, another participant could be observed searching the right option in the restaurant finder application. The task was to find a restaurant in Cologne but the participant could not find the right option. He comments this in the interview as following: *"I did not really find the restaurant option on first sight."* After a short period of searching and clicking different options the participant finds the right option and can continue to complete the task. This searching behavior did also indicate that the participant was developing a new mental model of the application through trial and error and therefore this incident can be classified as knowledge-based.

Another participant did also experience problems solving the same task. He did not find an option to change the city of interest. This option gets only visible through clicking into the search field but is not visible from the beginning. He comments in the retrospective interview that *"somehow there is no option for another*

city". He searches the application and clicks some wrong options and comments in the interview that he did not really understand what he was doing but then clicks on the search bar and finds the right option. This behavior also indicates a trial and error strategy to solve the problem and therefore we would classify this incident as knowledge-based.

A similar incident could be observed among a participant working with the radio application. The task was to find a way to mark the radio station she has found in a prior task to simplify the retrieving of this station. The participant formulated the goal that she wanted to add the station to list of favorites. She searched for an option to achieve this goal in the search-results window and comments this in the retrospective interview: *"Here I searched for an option to add the station to the favorites but then I realized that this was not possible in this menu."* The participant started subsequently to search the application for other ways and finds an option on the station page. This behavior indicates that the first plan of the participant fails and she then tries to develop another plan through trial and error until she achieves the formulated goal.

The following incident that could be observed does also illustrate the trial and error behavior that could be found with incidents associated with the knowledge-based level. The participant was working with the Twitter application and was trying to find an option to tweet but was struggling to find one. The participant is searching through the application and tries a button labeled "Direct Messages". The participant comments this in the retrospective interview as following: *"I thought that this could not be the right option. I tried it anyway because I did not find another way."* This quote illustrates that the participant had no plan of action to follow but tried to develop a new plan through trial and error to achieve his goal.

Working with the radio application five participants experienced problems solving the task to stop the station from playing. Most users did not find the "stop" button because it was not self-explanatory, due it is placed in a corner of the application not related to other playback options and has nearly the same color as the background. Some participants verbalized their helplessness in the retrospective interview, for example. *"I really found no way to stop the music"*. Most participants did find the button by chance, which the following quote indicates: *"I only wanted to try what this button does but I did not interpret it as a stop button"*. None of those five participants could be observed to follow a real plan but all searched the application, clicking buttons randomly, to find an option to stop the music. This behavior illustrates that the participants developed a new plan through trial and error and therefore we can classify these incidents as knowledge-based.

We could also observe incidents that can be related to knowledge-based behavior although the participants did not really comment an error or wrong behavior. In total, ten incidents could be observed in which all participant performed the right actions to fulfill the task, but doubted whether the task had been completed due to lacking feedback. All participants started to search for an option to revise if the action they conducted brought about the result they had intended. This search did happen in a trial and error manner and was not always successful for the participants. One participant comments this incident in the retrospective interview with the following statement: *"I wanted to find the list of favorites to check if this heart I clicked worked but I couldn't find it anywhere"*. Another participant experiencing the same problem comments: *"I could not distinguish if I really activated it or not and therefore I thought that it did not work"*. Also the Twitter application leaks a form of feedback if the tweet was sent or not. *"I was not sure if the tweet I composed was really sent",*

comments one participant and starts searching for an option to review if he completed his goal. Even if the participants completed the goal we classify this incidents on knowledge-based relying on the trial and error behavior resulting through the uncertainty.

Other Incidents

Some incidents that we could observe during the experiment cannot be classified confidently into the SKR-framework of Rasmussen (1983). The information provided through the analysis of the data related to the incident is not sufficient and therefore no explicit classification can be formulated.

An incident illustrating this leak of information could be observed among a participant working with the weather application. The participant experienced problems with the task to gather information about the weather forecast of the next day. She could be observed clicking on some blue dots indicating the page number of the forecast. This incident was commented through the participant with the following statement in the retrospective interview: *"I don't know if I clicked wrong there or it was not registered through the application."* If the participant had executed a shoving gesture instead of clicking the app would have displayed the weather information for the following day. Subsequent to this incident the participant clicks on another tab where the forecast for the next day is also presented. If the intention of the participant was to execute a shoving gesture we could conclude that this incident can be classified as skill-based. However, we leak enough information to formulate the intention of the participant. The next assumption that could be made is that the participant thought, "If I click on this dots, then the page will turn." But it is also possible that she only clicked on the dots in trial and error manner to develop a plan

how to achieve her goal. All of these options are possible and therefore this incident is classified as ambitious.

Participants could be observed commenting on problems they see in the applications but executing actions that lead to the completions of the task. For example a participant working with the weather application and trying to find information about the weather forecast comments: *"I'm confused through the indicator wheel on the left side and think that this cannot be the date."* Subsequent to this comment the participants turned the page of the application and found the information he was looking for. We only can assume that the participant was searching for a plan how to display the searched information and only found it through chance. If this was the case and the intention of the participant, this incident could be classified as knowledge-based but the information is not sufficient for such a categorization.

Working with the restaurant finder another participant could be observed hesitating to work on after clicking on the search option. The participant comments this in the retrospective interview as following: *"I was confused because I thought I would get a broader choice of options."* The participant found a restaurant in the prior task through the "Nearby" function, which gives the choice between restaurants and other points of interest, in contrast to the search function he was using during this incident. After a short period of time the participant works on and completes the task with no further problems. It could be formulated that the participant thought, "If I click on the search function then I will get a choice of options", but the participant gets not influenced through this assumptions or behaves in that way. Therefore we only possess over weak information to classify this incident and we can only assume that this incident is rule-based.

Roundup of the Qualitative Analysis

In the above analysis we were able to confidently classify 72 (79.9%) of the 111 observed incidents into the SRK-framework. Out of these incidents we classified 17 incidents as skill-based, 9 as rule-based and 46 as knowledge-based. The remaining 39 observed incidents (20.1%) could not be classified due to deficient information, or because the incidents only represent comments of the participants.

This shows that in the present case for roughly the crucial information for classification of four fifth of the incidents was available in the collected data. For the rest of the incidents the information was either not sufficient to make an explicit classification or the incident was not classifiable into the SKR-framework as discussed above. These results illustrate that the information provided through the here used data gathering technique is sufficient for classification of a vast amount of the observed incidents - although there is room for improvement. These improvements, which will be discussed later, provide the possibility to generate an optimal basis for the classification into the SKR-framework.

Since the raters did not remotely obtain a similar degree of accuracy in their classification it seems reasonable to assume that most of the valuable information was lost during the process of matching and presenting the resulting usability problems to the raters.

However, what kind of valuable information is lost in this process remains unclear. In the following, we therefore compare the information available to the raters during our study and the information available to us in the detailed qualitative analysis described above, and set this in relation to the resulting classifications made by the raters and ourselves. This comparison illustrates which information is crucial for the classification and was not available by the rater.

Since the raters only classified the matched usability problems and not, like we did above, the constituent usability incidents, a direct comparison of the resulting classifications is not possible. Therefore a comparison is made between the classifications of the constituent incidents that form the basis of the matched usability problems classified through the raters. This comparison shown in Table 5 illustrates that only in two cases the classification of the raters is in accordance with the classification of the qualitative analysis. In all other cases we can observe inconsistency among the classifications.

Table 6. Classification of the usability problems of the raters compared to the classification resulting out of the qualitative analysis.

Problem ID	Rater A	Rater B	Rater C	Qualitative Analysis
#1	SB	KB	SB	KB
#2	SB	KB	SB	KB
#3	KB	KB	RB	KB
#4	KB	SB	SB	SB
#5	KB	RB	RB	RB
#6	SB	KB	SB	KB
#7	KB	RB	SB	KB
#8	KB	KB	SB	KB
#9	KB	KB	RB	KB
#10	KB	KB	RB	KB
#11	SB	KB	SB	KB
#12	SB	SB	SB	RB
#13	RB	RB	RB	RB
#14	SB	RB	RB	RB
#15	KB	KB	KB	KB

To investigate this inconsistency, a comparison of the information available to the rater and gathered through the qualitative analysis has been made.

The following examples illustrate the method of comparison that has been executed for all usability incidents. The observations made in these two examples are representative of the findings observed in other comparisons.

The first two usability problems reported in the catalogue were related to the observation that participants had problems to find different features in the application. With this brief description alone, this problem might be interpreted the way that the participants could not execute the plan to use these features and therefore the raters suggest that these problems were skill-based. Viewing this problems in the context that the participant were searching for this features in a trial and error manner, and therefore, did not really have a plan of action, these problems are to be classified as knowledge-based. However this context was not available to the raters.

A second example to illustrate this lack of information can be observed with usability problem #12. All raters classified it as skill-based, relying on the information that the user fails to change the current location in the weather application by dragging the marker representing the location in a map. We can assume that the raters thought that the plan of the participant was to change the location of the weather application using this map but that this did not work out as he planned. Through the qualitative analysis the assumption of the participant that, “if a map with a marker is accessible then I can drag this marker to change the location”, was made accessible. Through this additional information illustrating the real intention of the participant we know that the problem reported was only the result of a wrong if/then assumption as discussed earlier and therefore this problem is related to rule-based behavior.

This comparison results in a confirmation of the earlier stated assumption that through the application of Hornbæk’s technique valuable information, regarding the intentions of the participant was filtered out. In addition to that, further information about the context of the usability problem, for example, which task the participant was trying to fulfill, was not accessible for the rater. Through the lack of this

information it was difficult for the raters to view the given problems in the context of the actual actions executed by the participants.

Summing up we can conclude from the above comparison that the context of the usability problems and the intention of the participant provide valuable information that is not accessible to the raters from the problem descriptions they had at their disposal. As illustrated earlier in the qualitative analysis it is important to know the intention of the user and the assumptions the user makes regarding the applications. Solely reporting characteristics of the problem, as in the Hornbæk method, seems insufficient to understand the underlying cause of the problem in terms of the SKR-framework.

Conclusion

The qualitative analysis of the usability incidents has shown that only with access to detailed data a reliable classification could be made. The method used to match and report usability incidents, filtered parts of this valuable information, which complicates the classification through the raters. Descriptions of usability problems, which are the result of the matching and reporting process, do not entail the intention of a participant or of assumptions the participants based their behavior on. The intention of the participant can best be accessed through a think-aloud method. However, the qualitative analysis also has shown that in some cases the intention of the participant was only accessible through the combination of the interview with observations of the actual behavior of the participant.

The intention of the participant is crucial for the classification of usability incidents into the SRK-framework. Therefore protocols for further investigation of usability problems should be designed to access this kind of information. We would suggest the retrospective think-aloud method for this purpose, because of different advantages also formulated earlier through Nielsen & Yssing (2004): The researcher possesses an uninterrupted recording of the interaction of the participant with the reviewed product and additionally the researcher has the possibility to conduct a more sophisticated interview with the participant through the possibility to stop the replay of the recording in the retrospective part. Through this method it is possible to access more information about the motivation of the participant for a certain action. In addition to that method, we suggest the development of an interview guideline used in this method. Specifically, the interviewer should ask the participant to every incident that could be observed in executive part why he or she conducted the specific actions and on which assumptions the participant based his decisions, to increase the

information regarding the intention of the participant. Supporting this method it would be useful to mark all incidents the researcher observes during this executive part on the recording to ensure that intentional information will be obtained for all incidents.

In addition to the method of gathering relevant data, the method of reporting usability problems should also be considered a crucial issue. This is especially the case in experiments were inter-rater agreement is compared. The report structure Lavery and colleagues (1997) and Hornbæk and Frøkjær (2008) suggest, does have the benefit that the usability evaluation's reliability increases through a structured report method, but it also lacks the opportunity to obtain more qualitative information about the occurrence of the usability problem. We would suggest adding such a category to this report method to include quotes or other observations into the report. Further, information about the task the participant was working on would provide a more detailed image of the goal the participant was trying to fulfill. Another option to achieve a more detailed image of the nature of usability problems and the context of their occurrence, video and audio recordings could be added to the report. These recordings would provide the raters with a more detailed image of the problems.

Even if this study did not provide the desired reliable basis to classify usability problems, our results stress the need for such a method. The low degree of agreement in our classifications shows that such a classification of usability problems is not trivial and a reliable method has to be developed. Furthermore, these results stress that if a classification was conducted during a study, the exact method of how this classification was conducted needs to be reported to ensure compatibility across studies. The classification of usability problems varied greatly in our experiments, even after provision of detailed instructions. It is likely that such inconsistencies were

present in earlier studies using classifications of usability problems, and, as mentioned in the beginning of this paper, it is possibile that these inconsitencies pose a threat for the reliablity and hence the validity of all conlcusions based on these classifications.

Reference

Besnard, D., & Cacitti, L. (2005). Interface changes causing accidents. An empirical study of negative transfer. *International Journal of Human-Computer Studies*, *62*(1), 105–125. doi:10.1016/j.ijhcs.2004.08.002

Courage, C., Jain, J., & Rosenbaum, S. (2009). Best practices in longitudinal research. *Proceedings of the 27th international conference on Human factors in computing systems extended abstracts - CHI EA '09* (p. 4791). New York, New York, USA: ACM Press. doi:10.1145/1520340.1520742

Embrey, D., & Lane, H. (2005). Understanding Human Behaviour and Error. *Human Reliability Associates*, *1*, 1–10.

European Committee for Electrotechnical Standardization Medical devices - Application of usability engineering to medical devices (2008).

Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology*, *21*(2), 137–143.

Gorlenko, L., & Englefied, P. (2006). Usability error classification. *CHI '06 extended abstracts on Human factors in computing systems - CHI EA '06* (p. 803). New York, New York, USA: ACM Press. doi:10.1145/1125451.1125610

Hertzum, M., Hansen, K., & Andersen, H. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165–181. doi:10.1080/01449290701773842

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. doi:10.1016/j.ijhcs.2005.06.002

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, *20*(6), 505–514. doi:10.1016/j.intcom.2008.08.005

Kjeldskov, J., Skov, M. B., & Stage, J. (2005). Does time heal?: a longitudinal study of usability (pp. 1–10). Canberra, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia. Retrieved from http://portal.acm.org/citation.cfm?id=1108368.1108394&coll=ACM&dl=ACM&CFID=53150963&CFTOKEN=10336642

Lavery, D., Cockton, G., & Atkinson, M. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information …*, *16*(4-5), 246–266. doi:10.1080/014492997119824

Nielsen, J., & Yssing, C. (2004). What Kind of Information does an HCI expert want? Retrieved from http://ir.lib.cbs.dk/handle/10398/6465

Norgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice?: an explorative study of think-aloud testing. *... Interactive Systems: Proceedings of the 6 th ...*, 209–218. Retrieved from https://blog.itu.dk/DIND-F2011/files/2011/02/norgaard_dis_2006.pdf

Norman, D. (1981). Categorization of action slips. *Psychological Review; Psychological Review*, *88*. Retrieved from http://psycnet.apa.org/journals/rev/88/1/1/

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *... , Man and Cybernetics, IEEE Transactions on*, (3), 257–266. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6313160

Reason, J. (1990). *Human error*. Retrieved from http://books.google.com/books?hl=en&lr=&id=WJL8NZc8lZ8C&oi=fnd&pg=PR9&dq=Human+Error&ots=AkRg5b8i_g&sig=G3fMgqRcXKYa4JsLNOSnd6Wqej4

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18839484

Zapf, D., Brodbeck, F., & Frese, M. (1992). Errors in Working with Office Computers: A First Validation of a Taxonomy for Observed Errors in a Field Setting. *International Journal of ...*. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10447319209526046

Appendix A - Usability test consent form

Please read and sign this form.

In this usability test:

- You will be asked to perform certain tasks with different applications.
- You will be asked to fill in a questionnaire.
- You will also be asked to comment your actions afterwards.

Participation in this usability study is voluntary. All information will remain strictly confidential. The descriptions and findings may be used to help improve the applications. However, at no time will your name or any other means of identification be used. You can withdraw your consent to the experiment and stop participation at any time.

If you have any questions after today, please contact me Adrian Haar (a.haar@student.utwente.nl).

I have read and understood the information on this form and had all my questions answered. I agree that I will be videotaped in this study and I understand that the information and videotape are for research purposes only and that my name and image will not be used for any other purpose.

_____ _____

Participant's Signature Date

_____ _____

Researcher Date

Appendix B - Experiment 1 Instruction for coding usability problems

Thank you for participating in my study and for helping me to get inter rater scores for my algorithm to categorize usability problems.

This document consist out of two parts:

- The algorithm
- The usability error catalogue

  The problems described in this catalogue are reported in the study of different personal-management apps. Therefore it is possible that some problems are contradictory.

Please use the algorithm to categorize the given problems into the three categories skill-based, rule-based and knowledge-based.

Note your categorization like the following table and report it back to me:
Please use the following codes for the categorization
1 = skill-based, 2 = rule-based, 3 = knowledge-based

Problem Number	Categorization
#1	
#2	
#3	
#4	
#5	
#6	
#7	
#8	
#9	
#10	
#11	
#12	

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
1	**Description:** A button next to the input field with a plus on it suggest to add the input as a new to do item, but does not have a function.	**Problem Location:** Input Bar on the top of each screen
	Outcomes and Risks: Users may misinterpret the button as a enter button and get frustrated by the “malfunction” of this button and stop to search for another way to add the item.	**Design Recommendations:** Remove the button and add a enter icon to signal the user that the enter key should be user to add items
2	**Description:** The field that is intended to mark if a task is done or not gets confused with a field to mark tasks to edit settings	**Problem Location:** Field on each task
	Outcomes and Risks: Users may accidently mark tasks as done in the intention to edit settings like due date or during a try to search a way to move multiple items into a folder/list.	**Design Recommendations:** Mark the existing fields with “done”

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
3	**Description:** Some task “disappear” -> If tasks are marked as done in a list or in the inbox and the users navigate to a different menu and then come back the tasks “disappeared”	**Problem Location:**
	Outcomes and Risks: Users may misinterpret the disappearance as the task got deleted by accident and do not realize that the completed task are only moved to the “done” folder	**Design Recommendations:** Some sort of feedback to inform the users that the task did not disappear
4	**Description:** Users hat the problem that the misspelled their password wrong	**Problem Location:** Log In Screen
	Outcomes and Risks: Through small keyboards on smartphones or even tablet computers, tipping your password is not always easy and can get frustrating when you misspell it frequently.	**Design Recommendations:** Give an option to show the password in save locations.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
5	**Description:** Users confuse creating a list with creating a task.	**Problem Location:** Main Menu
	Outcomes and Risks: To create new items, users have to navigate into the inbox or into a list. If they stay in the main window they only create a new list, which share the appearance of a single item. Therefore it is hard to distinguish for the users if they created a list or a single item.	**Design Recommendations:** List should be different in their appearance to single to do items to communicate visually the difference between them. In Addition to that a list could have the word “List” on it to stress their function.
6	**Description:** The user tries to edit the task in the main menu	**Problem Location:** Main Menu
	Outcomes and Risks: Users do not find the options to edit a to do item directly and search the main menu for more options. The right way to edit the item would be to hold the item until more options would come visible.	**Design Recommendations:** Some hint in the background would be helpful e.g. “Hold to edit”

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
7	**Description:** Users get confused by the difference between due date/reminder	**Problem Location:** Menu to edit the properties of a item
	Outcomes and Risks: Users do not really get what the difference is between the due date and the reminder function. Both options asks the user to enter a date but do not explain what the difference is.	**Design Recommendations:** More explanation could be useful. Instead of simple labeling, questions could do a better job e.g. “When is this item due?”/ “When do you want to be reminded for this item?”
8	**Description:** After completing editing a task the user is confused if the changes he made are really saved.	**Problem Location:**
	Outcomes and Risks: Users often did the right thing to edit a task but were confused if this action was right because they got no feedback if e.g. the reminder is saved and if they will be reminded on the given date.	**Design Recommendations:** Some little messages as feedback would help the user.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
9	**Description:** Users do not find a way to mark a task as “done”	**Problem Location:**
	Outcomes and Risks: The app intends that task has to be stroked out with a gesture. This becomes a problem because there’s no visual hint for the user to figure this out.	**Design Recommendations:**
10	**Description:** Users quit the app by accident.	**Problem Location:** Main Menu
	Outcomes and Risks: Users hit the “go back” button one time more as needed, quit the app and go back to the home screen.	**Design Recommendations:** If the user tries to quit the app a conformation prompt could solve the problem e.g. : Do you really want to close the app?”

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
11	**Description:** Users confuse the “Add Bar” with a search bar and don’t use it to add new task	**Problem Location:** Input Bar on the top of each screen
	Outcomes and Risks:	**Design Recommendations:** Put the bar on the bottom of the screen to make it harder to confuse with a search bar.
12	**Description:** Vague / ambiguous icons	**Problem Location:** Sub-menu of each item
	Outcomes and Risks: Users can not really anticipate the function behind each icon because they are not meaningful enough	**Design Recommendations:** Redesign the icons to make them more meaningful

Appendix C – Experiment 2 Instruction for coding usability problems

Thank you for participating in my study and for helping me to get inter rater scores for my algorithm to categorize usability problems.

This document consist out of two parts:

- The algorithm and a instruction text how to work with it
- The usability error catalogue

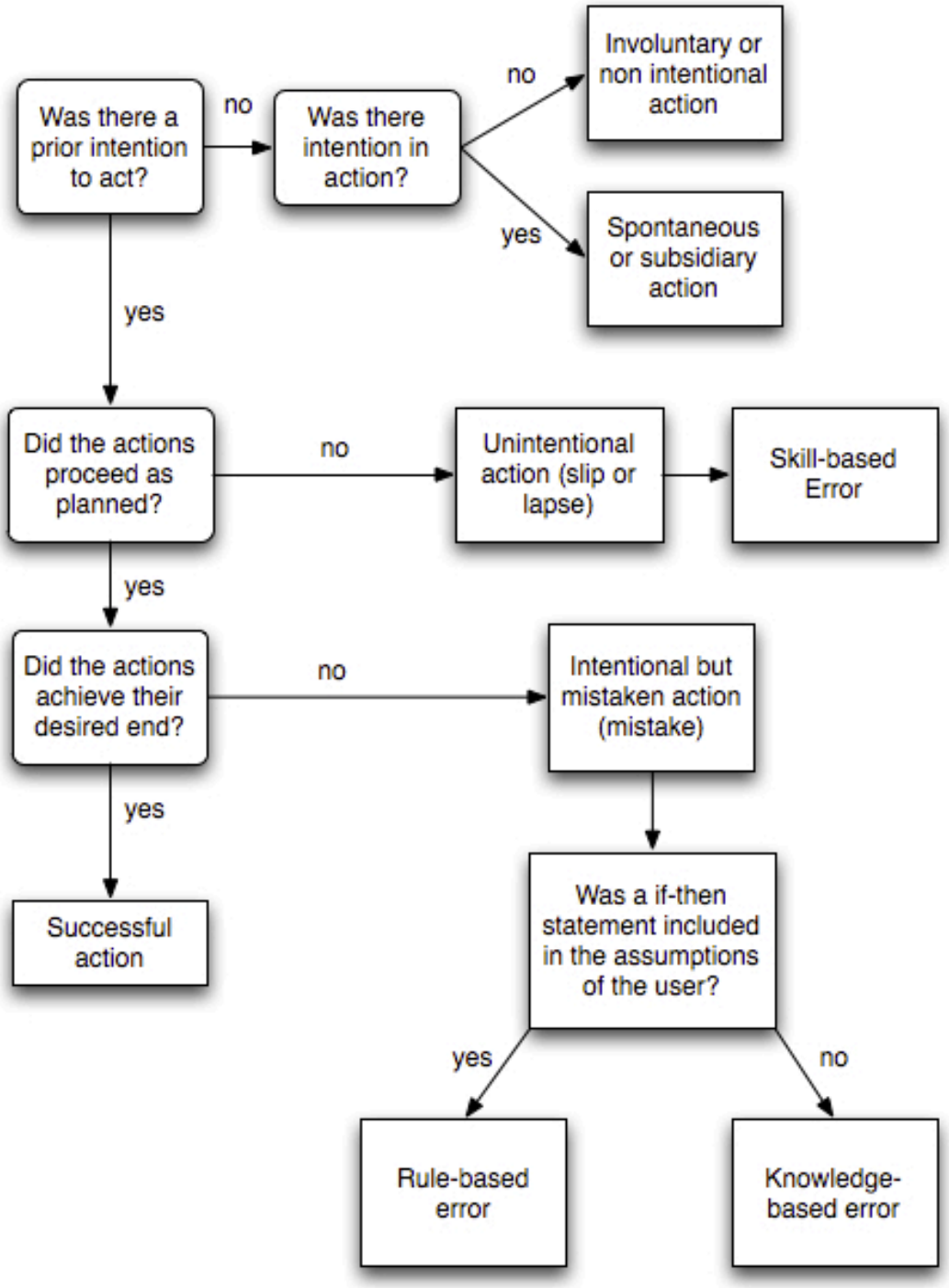
  The problems described in this catalogue are reported in the study of different personal-management apps. Therefore it is possible that some problems are contradictory.

Please use the algorithm to categorize the given problems into the three categories skill-based, rule-based and knowledge-based.

Note your categorization like the following table and report it back to me:
Please use the following codes for the categorization
1 = skill-based, 2 = rule-based, 3 = knowledge-based

Problem Number	Categorization
#1	
#2	
#3	
#4	
#5	
#6	
#7	
#8	
#9	
#10	
#11	
#12	

The following page shows the algorithm you will use to categorize the given usability problems into the three categories. On the following pages we will provide you with an introduction how to use this algorithm.

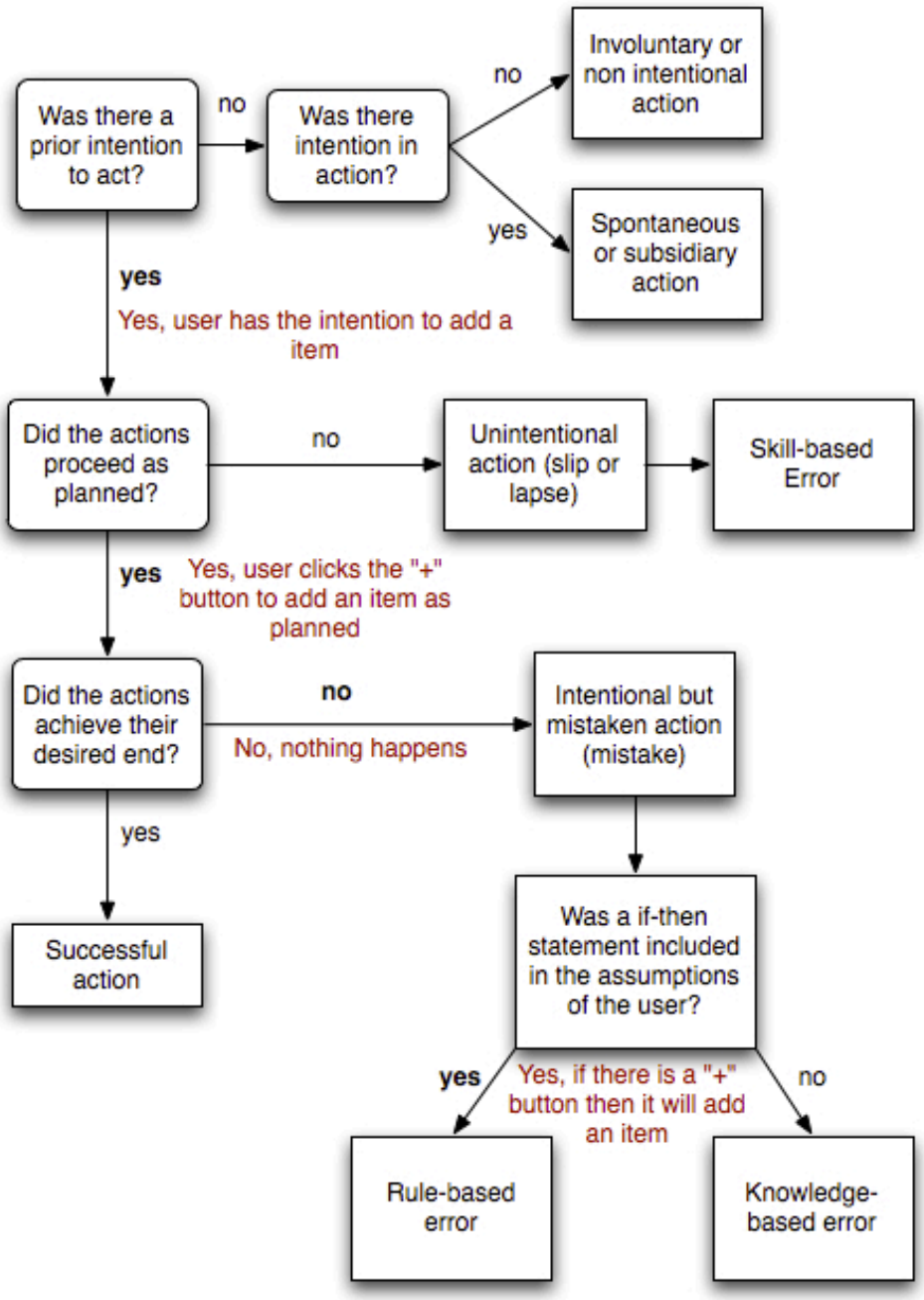


How to use the algorithm:

1. Read the problem description for every usability problem and try to get a good image of the problem and about what the participant thought during the experiment.
2. Start at the upper left corner of the algorithm and try to answer the question in matters of the problem you are categorizing.
   Was there a prior intention to act? – Generally speaking you will answer this question with "yes" because all the participants had a task to solve and therefore an intention to act.
3. The next question is: "Did the actions proceed as planned?" If users have the intention to solve a task they will have a plan how to work on this task. If they cannot complete or execute this plan out of some reasons we speak about a slip or a lapse and you have to answer this question with "no".
4. If the user completed the plan he chose but he did not reach the desired end state we can talk about a mistake. In this case the plan the user had was wrong or not applicable for the situation.
5. The last decision point of the algorithm helps you to distinguish between a rule-based and a knowledge-based mistake. Usually users try to solve a task/problem with stored IF/THEN assumption if the "pattern" of the problem is familiar. An example could be that the user tries to stop the music and thinks: "IF the button is red, THEN it will stop the music". In that case, if the red button will not stop the music, we talk about a rule-based mistake. A rule-based mistake occurs if either the user did work with a wrong rule or the rule was not applicable for this program.
6. If the user does not work with an IF/THEN assumption we would talk about a knowledge-based error. This occurs when the user is not familiar with the given situation and do not have an applicable IF/THEN rule.

To illustrate this workflow we will provide you with an example.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
1	**Description:** A button next to the input field with a plus on it suggest to add the input as a new to do item, but does not have a function.	**Problem Location:** Input Bar on the top of each screen
	Outcomes and Risks: Users may misinterpret the button as a enter button and get frustrated by the "malfunction" of this button and stop to search for another way to add the item.	**Design Recommendations:** Remove the button and add a enter icon to signal the user that the enter key should be user to add items


Was there a prior intention to act?
no
Was there intention in action?
no
Involuntary or non intentional action
yes
Spontaneous or subsidiary action
yes
Yes, user has the intention to add a item
Did the actions proceed as planned?
no
Unintentional action (slip or lapse)
Skill-based Error
yes
Yes, user clicks the "+" button to add an item as planned
Did the actions achieve their desired end?
no
No, nothing happens
Intentional but mistaken action (mistake)
yes
Successful action
Was a if-then statement included in the assumptions of the user?
yes
Yes, if there is a "+" button then it will add an item
no
Rule-based error
Knowledge-based error

Problem Nr.	Description	Problem Location (Application)
	Outcomes and Risks	Design Recommendations
1	**Description:** The magnifying glass, representing the search function is not easy to find because it has the same color as the background	**Problem Location:** (TuneIn) In the upper right corner of each menu
	Outcomes and Risks: Users may not find the search function, with the risk to not find their desired radio they're searching for.	**Design Recommendations:** Give the magnification glass another color or give a description in text form next to it, make it easier to find.
2	**Description:** The Stop button, to stop the radio, was not easy to find for some users.	**Problem Location:** (TuneIn) The Stop button is a square in the button of the screen in the same shade of the background available in each menu.
	Outcomes and Risks: Users did not find a way to stop the radio	**Design Recommendations:** The square for stopping music is normally a wide used symbol but does only sense in the context with other buttons representing functions to control the music. Therefore the button should be used in this context or given a textual explanation.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
3	**Description:** Users find the option to assign a station as a favorite, but are not sure if it really worked because the are missing the feedback	**Problem Location:** (TuneIn) In the "Station screen" the heart button in the upper right corner
	Outcomes and Risks: Users may click to more times as needed on the favorite button because the are not sure if it worked and therefore the risk exist that they "unfavorite" the station again. Further the risk exists that users will ignore this feature because they think that the feature does not work.	**Design Recommendations:** Some sort of feedback to inform the users that the station is added to the favorites and maybe even a hint where the can find the favorites.
4	**Description:** Users tend to use the back button to often and therefore close the application they use accidentally.	**Problem Location:** General problem with every application
	Outcomes and Risks: This problem could be general disturbance for the ease of use of all applications if the user has to start the application again, further it could be a risk that data would not be saved through this accidentally closing of the application	**Design Recommendations:** The application could ask the user for a conformation if he really wants to close the application.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
5	**Description:** The application shows in the overview of all restaurants only the name of the district of the searched city and not the name of the city itself.	**Problem Location:** (Yelp) Restaurant overview after a search
	Outcomes and Risks: Users could be confused if the search was successful or not if they are not familiar with the different districts in the searched city	**Design Recommendations:** Add the name of the city to the details of a restaurant in the search overview
6	**Description:** Users could not find the option to change the city where they want to find a restaurant, because the only option appears as the click the search bar. As they click the search bar a second search bar appears where the user could search in a specific city.	**Problem Location:** (Yelp) see Description
	Outcomes and Risks: The outcome of this problem is that some users needed a long time to find a way to search for a restaurant in a different location. Most users found this option only coincidently by clicking the search bar.	**Design Recommendations:** A solution for this problem could be to make the second search bar for the location permanent and locate it next to the default search bar.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
7	**Description:** Users needed a lot time to find an option to edit their profile. The “edit profile” button is placed in a submenu of the category “Me” but most users anticipated to find this option right in the category and not in a submenu.	**Problem Location:** (Twitter) see Description
	Outcomes and Risks: Users can get demotivated and frustrated while searching for the right option to edit their profile	**Design Recommendations:** Place the button into the Main Menu of the category “Me” and not in a submenu
8	**Description:** Some users searched a long time for an option to tweet, besides that the button with this functions is positioned in the upper right corner in every menu.	**Problem Location:** (Twitter) see Description
	Outcomes and Risks: To tweet is one of the basic functions of a twitter client and if a user cannot find it he or she could get very frustrated.	**Design Recommendations:** Some users reported that the icon in the right corner did not have enough meaning for them, therefore a hint in text form could help (e.g. “Tweet”)

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
9	**Description:** If users submitted a tweet they were not sure if this really was send or not, because they had to refresh their newsfeed manually to see their tweet	**Problem Location:** (Twitter) see Description
	Outcomes and Risks: If users are not sure if their tweet was send they could tend to send it again without knowing that it worked the first time	**Design Recommendations:** Give the user feedback (e.g.: "Your tweet was send successfully) and/or refresh the newsfeed automatically.
10	**Description:** As a user changed the name of his profile, some users were not sure if it really worked because own tweets that are shown in the "Me" category had still the old name. Only as they clicked on another category they and back to "Me" they could see that the name really changed.	**Problem Location:** (Twitter) Category "Me"
	Outcomes and Risks: If users are not sure if their change was already applied they could tend to change it again without knowing that it worked the first time	**Design Recommendations:** The tweets in the "Me" category should have the new name already and/or give the user some feedback (e.g. Your new name is now "…")

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
11	**Description:** Users could not change the given temperature from Fahrenheit to Celsius. Most of the users found the right option in the settings menu but could not interpret the given information. The only option to change the temperature is to change all units from Imperial to Metric. Most users did not know that this would change the temperature form Fahrenheit to Celsius.	**Problem Location:** (AccuWeather) Settings Menu
	Outcomes and Risks: Even if users found this option, they did not choose metric because it was not clear that it would change the unit of temperature. Therefore some users got really frustrated because they could not find a way to change the unit	**Design Recommendations:** In addition to the option to change all units von imperial to metric an option should be added to choose the unit of the given temperature. Some users tried to click the temperature self to change its unit, which also could be an option to implemented.
12	**Description:** Users try to change the location through searching a city on the weather app but nothing happens if the click the map	**Problem Location:** (AccuWeather) Weather map
	Outcomes and Risks: Users could get frustrated through the fact that nothing happens	**Design Recommendations:** Add a option to change the location over the map or give some feedback that its not possible to change the location if the try to click the map.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
13	**Description:** Users try to click the page indicator but nothing happens	**Problem Location:** (AccuWeather) Main Menu and Forecast Menu
	Outcomes and Risks: The user could get frustrated through the fact that the page indicator looks “clickable” but nothing happens.	**Design Recommendations:** Change the page as a user clicks on the page indicator. This would also emphasize the function of this indicators and the user would directly know what this numbers indicate.
14	**Description:** Some users get confused through the fact that the icons representing other menus are grey and assume that grey buttons cannot be pressed	**Problem Location:** (AccuWeather)
	Outcomes and Risks: Users could think that the application does not work because they cannot click another menu than they are.	**Design Recommendations:** Indicate in another way which menu is chosen through another choice of colors.

Problem Nr.	Description	Problem Location
	Outcomes and Risks	Design Recommendations
15	**Description:** Users cannot interpret the information the page indicator gives in the forecast menu. The page indicator only gives a number but for the user it's not clear which day is now shown in the forecast	**Problem Location:** (AccuWeather) Forecast Menu
	Outcomes and Risks: The user could get confused about the day he is getting information about and can misinterpret those.	**Design Recommendations:** Instead of giving only a number as indicator the weekday would be more specific.