Argumentation models for usability problem analysis
in individual and collaborative settings

Ebba Thora Hvannberg, University of Iceland

Effie L-C Law, University of Leicester

Gyda Halldorsdottir, University of Iceland

Corresponding author:

Ebba Thora Hvannberg

University of Iceland

Dunhaga 5

107 Reykjavik

Iceland

ebba@hi.is

+354 525 4702 +354 897 9196

To cite this article:

Ebba Thora Hvannberg, Effie L.-C. Law & Gyda Halldorsdottir (2018) Argumentation Models for Usability Problem Analysis in Individual and Collaborative Settings, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2018.1454142

Argumentation models for usability problem analysis in individual and collaborative settings

Abstract

Consolidating usability problems from problem lists from several users can be a cognitively demanding task for evaluators. It has been suggested that collaboration between evaluators can help this process. In an attempt to learn how evaluators make decisions in this process, we studied what justification evaluators give for extracting usability problems and their consolidation when working both individually and collaboratively. An experiment with eight novice usability evaluators was carried out where they extracted usability problems and consolidated them individually and then collaboratively. The data were analysed by using conventional content analysis and by creating argumentation models according to the Toulmin model. The results showed that during usability problem extraction novice usability evaluators could put forward warrants leading to clear claims when probed, but seldom added qualifiers or rebuttals. Novice usability evaluators could identify predefined criteria for a usability problem when probed and this could be acknowledged as a backing to warrants. In the individual settings, novice evaluators had difficulty in presenting claims and warrants for their decisions on consolidation. Although further study is needed, the results of the study indicated that collaborating pairs had a tendency to argue slightly better than individuals. Through the experiment novice evaluators' reasoning patterns during problem extraction and consolidation as well as during their assessment of severity and confidence could be identified.

Keywords: usability testing, user studies, empirical studies in HCI, argumentation, collaboration, consolidation.

Research highlights:

- Novice usability evaluators can well justify their decisions on usability problem extraction, giving warrants leading to clear claims. At least implicitly as part of an argument, backings to warrants are seen in half the cases as a reference to predefined criteria of what a usability problem is. When asked to state criteria explicitly evaluators could provide them. Evaluators rarely qualified their claims and seldom included rebuttals to claims.
- Previous research has shown that teams can argue better than individuals. The results of this study showed that in both the individual and the collaborative settings, novice usability evaluators had difficulty in giving justifications for their decisions on consolidation.
- Novice usability evaluators would benefit from using argumentation templates, including claims, warrants, backings, qualifiers and rebuttals to improve their arguments when extracting and consolidating usability problems.
- Novice usability evaluators backing arguments with own experience are better convinced of their decisions than when grounding decisions by referring to users.
- In their argumentation, novice usability evaluators cited criteria for a usability problem, severity definitions, users' behaviour, the context of a problem, video clips and their own experience. Novice usability evaluators reported that they consulted video clips in two thirds of the usability problems and could suggest how to improve the design in all cases.

1 Introduction

A usability problem (UP) is something that causes users difficulty in using a system. Operationally, a UP is identified as such if it meets one or more of predefined criteria (John & Mashyna, 1997). In usability testing users are observed while performing tasks and a list of UPs are extracted (Niels Ebbe Jacobsen, Morten Hertzum, & Bonnie E. John, 1998b). It is not only the number of UPs that matters, but also finding their causes in design and difficulties experienced by users (Cockton & Lavery, 1999). Applying usability evaluation methods appropriately is important for understanding and influencing the outcome of the overall process of a system design (Dumas & Redish, 1999, p. 25). Systematic studies have concluded that usability practices are not documented thoroughly (Boren & Ramey, 2000), which can vary across cultures (Clemmensen, 2011) and contexts (Furniss, 2008). To capture the structure of diverse usability evaluation methods, a systematic review has defined three general phases of usability evaluation: *capturing* involves collecting usability data; *analysing* involves extracting UPs; *critiquing* suggests improvements (Ivory & Hearst, 2001). A method for extracting UPs covering the two latter phases was suggested by Cockton and Lavery (1999) and the method was derived from basic concepts in usability evaluation. Accordingly, four different judgements, or steps, are involved in UP extraction, which can be viewed as a single-step decision. The first one determines the boundaries between episodes of interaction. The second one analyses relevant difficulties, where similar difficulties can be collected or merged into generalizations. The third adds causal analyses to difficulties and the fourth one recommends changes to remove causes. In this paper we have chosen to use the term ‘decision’ but what precedes decision includes making sense of the data and the situations of user interaction (Boland, 2008).

Involving multiple users in usability evaluation typically results in an unpredictable overlapping of UPs identified by these users. This requires usability evaluators to consolidate or merge the identified UPs that are duplicates of the same UP and to produce a UP master list (Law & Hvannberg, 2008). The consolidation process consists of two main steps: 1) Filtering duplicate UPs from a list of UPs identified either by a single user when performing a certain task within a system or by an analyst when inspecting it; 2) Merging unique combined UPs between different lists identified by multiple users/analysts (John & Mashyna, 1997). Research has shown that there is a clear need for effective consolidation of usability problems. In a survey, 50.8% of 147 software development practitioners found that the issue with usability problems is that they include duplicate reports (Yusop, Grundy, & Vasa, 2016).

According to Cockton and Lavery (1999) a collaborative causal analysis by a multi-disciplinary team is generally preferable to analysis by a single usability professional. Analysing causes collaboratively, once the difficulties have been analysed, is seen in many industrial settings (Cockton & Lavery, 1999). Dumas and Redish (1999, p. 334) pointed out that developing usable products would be a team effort, especially when discussing solutions to problems. Furthermore, they noted that usability engineers would engage in arguments with designers when supporting recommendations, requiring usability engineers to stand their ground backed up by data. Despite these issues were documented almost two decades ago, they remain relevant in the recent years (e.g. (Bornoe & Stage, 2017)), implying that more research needs to be done to understand them.

In the scientific community quality control is maintained through the process of argumentation (Kuhn, 1992), and arguments showing the appropriateness of evidence and validity of claims are the heart of any discourse. Theoretical and practical approaches to argumentation can share basic principles such as considering relevant facts, claims should be well-grounded; supporting and refuting the claims should be grounded. In general, such principles of argumentation are needed across a wide variety of fields (Scheuer, Loll, Pinkwart, & McLaren, 2010). Although not being formal in a mathematical sense, Toulmin (1982) made a useful schematic representation of an argument. Essentially, he made a model of an argument as a claim supported by data and backed by a warrant. The elements supporting a claim are data involving factual information as an interpretive standpoint, a warrant justifying the inference between the data and claim, and backing or strengthening the warrant (Stegmann, Wecker, Weinberger, & Fischer, 2012; Toulmin, 2003). Such argumentation models have wide applicability.

Modern industrialised design in, for instance, engineering and materials science utilises both scientific knowledge and intuitive thinking; this is generally referred to as scientific design (Cross, 2001). Argumentation is as applicable in design as in science. Dalsgaard, Dindler, and Fritsch (2013) used design argumentation to teach design with the Toulmin model. The purpose was to use design as a vehicle for exploring theory and method in teaching. In their study, Dalsgaard et al. (2013) asked students to ground design arguments. The results showed that argumentation worked as a way for students to express qualities and shortcomings of their design choice. Dalsgaard et al. (2013) identified three categories of grounding, namely that students may ground arguments in theoretical notions, empirical data and material with which they work.

Although studies have investigated argumentation in design (Dalsgaard et al., 2013; Nörgaard & Höegh, 2008), from a theoretical perspective, more research is needed to study to what extent designers use scientific methods in their work (Cross, 2001; Fischer, Lemke, McCall, & Morch, 1991). The overall aim of the research presented in this paper is to learn about the evidence and validity presented by novice usability evaluators by exploring argumentation patterns that they provide individually and in teams during different phases of usability evaluation, i.e. UP extraction, consolidation, and severity assessment. Given the requirement to investigate further the justification evaluators give for problem extraction, filtering and merging, we set out to study the topic by analysing evaluators' reasons qualitatively and by constructing argumentation models. The research questions put forward for the study were:

RQ1 What justifications and support do evaluators give for UP extraction, and how do they qualify their decisions?

RQ2 How do evaluators explain how confident they are of problems?

RQ3 How do evaluators consolidate UPs in an individual vs. a collaborative setting and how easy do they think it is?

RQ4 How do evaluators explain the severity ratings of problems extracted and how do they analyse the severity of consolidated UPs in a collaborative setting?

To answer these questions, we ran an experiment where we asked evaluators to extract usability problems, filter them individually and then consolidate them collaboratively. After each of these three phases we interviewed evaluators and asked why they had extracted, rated, filtered or merged problems.

2 Background

Analysing the output of usability evaluation is challenging, and surveys have shown that more tools are needed (Følstad, Law, & Hornbæk, 2012). Although, usability evaluation strategies are mainly task oriented with performance being a key indicator for UPs, user behaviours and verbalizations are accepted as separate data sources to complement such data (Følstad et al., 2012). Additionally, raw data from interviews with evaluators can be appropriate as an interpretive standpoint for analysis of an evaluation. Several studies have investigated this aspect, e.g. the study of Nørgaard and Hornbæk (2006) indicated that evaluators had strong ideas about usability problems of systems and Yusop et al. (2016) found that software developers justified the cause of the problem by citing their own knowledge.

Little is known about the consolidation procedure and its impact on the final outcome. In a systematic literature review of reporting usability defects, Yusop, Grundy, and Vasa (2017) found four studies that use usability defect data for identifying similar usability problems. Hornbæk and Frøkjær (2008) carried out a study examining the outcome of applying four different techniques to match problems. The finding of the study showed that matching by similarity of changes to remove a usability problem resulted in the lowest agreement among evaluators and that the matching technique that is applied can strongly affect the results. It has been suggested that consolidating usability problems is not easy, methods to guide practitioners are lacking, and that little is known on how consolidation is carried out in practice (Hornbæk, 2010).

Researchers have investigated whether individuals consolidate problems differently than teams. The results showed that collaborative merging decreased the number of absolute UPs but that the severity ratings of UPs increased substantially (Law & Hvannberg, 2008). A finding of Følstad et al. (2012), who conducted a survey with usability evaluators, was that collaboration between practitioners is important, not least to improve the reliability of UP extraction. While group discussion is recommended in the literature, group analysis of usability testing data is seldom emphasised (Følstad et al., 2012). The issue of collaboration has also been investigated in the discipline of software engineering where e.g. face-to-face team meetings on defect inspection did not find significantly more defects than when inspectors worked individually. However, the meeting-based review method was significantly better at reducing the level of false positives (Johnson & Tjahjono, 1998). Sauer, Jeffery, Land, and Yetton (2000) set forth a behavioural theory of group performance describing several factors influencing the effectiveness of software reviews. Their research showed that the expertise of individual reviewers is the most important factor in determining the effectiveness of software reviews. The process suggested by Sauer et al. (2000) is to have individual reviewers inspect software, followed by having a single collector that separates confirmed true defects from the output of individual reviewers, and the final step suggested is to have a pair of experts review the outcome and look for false positives among those defects identified by just one reviewer. Although in the above processes of software review there is mentioning of finding the same or true defects there is no description of how experts perform this task. While we have drawn parallels of the processes of usability evaluation and software reviews there are differences between the two, mainly in the capturing stage where data on users' interaction with a product is captured in the former case whereas an inspection on software code, design or product is made in the latter case.

Studies of rational behaviour and selfishness have shown that teams tend to be more rational than individuals (Kugler, Kausel, & Kocher, 2012; Maciejovsky, Sutter, Budescu, & Bernau, 2013). Such studies typically involve playing a type of game and rely on four assumptions. One of them is that people know their interest and preferences. The second assumption is that people can determine what actions would best serve these interests and, third, that these interests provide material payoff. The fourth assumption is that everyone knows the rules of the game and are rational (Kugler et al., 2012). Research results have shown that exchanges of arguments in dyads or teams, are meant to convince and exceed individual reasoning (Mercier, 2016). Previous research has shown that having collaborators explicitly state their own findings before collaborating, compared to implicitly deriving them, would help their argumentation in the collaborative setting (Papadopoulos, Demetriadis, & Weinberger, 2013).

Design rationale schemes have been suggested to help designers to justify their decisions during design. Rittel and Webber (1973) viewed design as a process of negotiation and deliberation (Moran & Carroll, 1996) and while Rittel (1987) recognised that much of the mental activity resides and occurs in the subconscious, he proposed design as a process of argumentation. Several representations of design rationale have been suggested. Shum (1996b) gave an overview of argumentation techniques and case studies showing how argumentation-based approaches had been used in software development. He raised issues such as cost-benefit trade-offs of applying design rationale representations and the necessity of training designers in the techniques. Some of these representations were QOC (Question Option Criteria) (MacLean, Young, Bellotti, & Moran, 1991) and claims analysis (Carroll & Rosson, 1992). The usability of such representations was studied (Shum, 1996a) and how design rationale was used in meetings (Olson et al., 1996).

Claims are seen as weighing positive and negative aspects of design options. Sutcliffe and Carroll (1999) developed frameworks so that claims would be reusable. They also developed schema for claims that included a theoretical backing. This was seen as intended more for academic purposes than practical ones. Blandford, Keith, and Fields (2006) found that claims analysis was more difficult to learn, communicate to developers and to apply effectively than expected. For example, it proved difficult for developers to work out the positive and negative consequences (Blandford et al., 2006, p. 21). But developers found it much easier to identify positive effects of their own designs. They described positive features with more hope than confidence and did not back up their hope by citing theories or empirical studies (Blandford et al., 2006, p. 22).

Arguments can be modelled as diagrams in various ways (see e.g. Reed, Walton, and Macagno (2007)). Figure 1 shows a model of an argument using Toulmin theory. Data are the foundation of a claim. A warrant is the operational name used for the argument of moving accepted data to a claim, i.e. warrant is a set of preconditions accepted data must fulfil to become a claim. If present, backing serves as credentials used to certify an argument expressed in a warrant. Optionally, a qualifier is supposed to register the degree of a force of the claim. A tenable claim depends on the absence of a rebuttal, i.e. an appended rule defeating a warranted conclusion (Toulmin, 2003). Examples of backing include evidence such as statistics or expert opinions in line with the warrant. Rebuttals point to circumstances where a claim cannot be true (Toulmin, 2003). Some research has chosen not to distinguish between the two components of the argumentation, backing and warrant (Stegmann et al., 2012).

Figure 1 Toulmin Theory: A model of an argument

As an example from our data analysis (Table 1), facts revealed that a search button was missing. The participant claimed that the particular UP could be confirmed as a usability problem. According to the participant, a warrant for this claim was a precondition that a missing search button was hindering a user in performing a particular task. We could see that the participant used the observational reports (see Figure 3) describing user's actions to form the warrant. In this particular example, the participant did not strengthen or weaken the claim with a qualifier and there was no rebuttal describing responses to the claim.

Table 1 An example from data analysis leading to a model of an argument

Apart from Toulmin's, several argumentation models exist, each with its advantages and disadvantages (Noroozi, Weinberger, Biemans, Mulder, & Chizari, 2012). Some of these models have been used in the area of usability research to see how persuasive usability feedback is (Nörgaard & Höegh, 2008). Using data from two empirical studies, Nörgaard and Höegh (2008) referenced both the Toulmin model and the Aristotle model (Aristotle, 2006) to understand how argumentation is structured in several formats of usability problems. Although Toulmin theory has many advantages, researchers investigating collaborative argumentation have had difficulties in analysing dialectical features of dialogical argumentation, using the model (Nielsen, 2013).

3 Research design

The research study included observing three processes: a process of individual UP extraction, individual filtering and collaborative merging. During UP extraction, a participant was asked to extract as many UPs as possible from narrative observational reports that had been given to him/her. Second, in the individual filtering, a participant was asked to filter out any duplicates and consolidate similar UPs. Finally, in the collaborative merging, two participants were required to consolidate their respective lists of UPs prepared in the individual sessions and make a master list of identified UPs. These processes of UP extraction and merging reflect usability evaluation activities and we chose to follow those practices rather than applying a controlled experiment. Hence, we did not try to control for carry over effects by exchanging the order of the activities done individually or collaboratively.

Several characteristics were observed as participants performed these processes. Table 2 shows an overview of the characteristics collected after each process and the types of outcomes from the analysis. First, we wanted to see if participants could describe the strategies they used for deciding how to extract problems. We used this data to formulate argument models and researched if these strategies could be categorised. We asked participants about their confidence in their decisions and if their decision of extracting a problem was related to their previous experience in using the system. For these two characteristics we were interested in knowing whether categories emerged. Deciding on the severity of an extracted problem is a complex activity (Hertzum, 2006) and hence we asked participants how they had decided the severity and used the data to formulate argument models and categories. When studying utility of redesigns vs. usability problems to developers, Hornbæk and Frøkjær (2005) found that developers valued redesigns more than usability problems. When investigating users' verbalisations during think-aloud testing, Hertzum, Borlund, and Kristoffersen (2015) found that redesign proposals were infrequent and suggested that they should be prompted for. Therefore, in this study we sought to see if redesign proposals formed visible categories.

Table 2 Characteristics collected and outcome of analysis during individual problem extraction

---

To understand how pairs consolidated problems differently than individuals, we described the characteristics and the types of outcomes of the analysis in **Table 3**. For both individual filtering and collaborative merging, our main interest was to learn what kinds of strategies participants used. For this, we analysed the data by creating argument models and forming categories. We were also curious to analyse how easy participants found it to make decisions. For pairs we examined how they decided on the severity of the consolidated problems. Since we felt this was a central factor, we derived categories of reasoning patterns for deciding on severity and extracted models of arguments from the qualitative data.

Table 3 Characteristics collected and outcome of analysis during individual filtering and collaborative merging

---

4 Methodology

In the following sections we describe the methodology of the study. In the first subsection, participants and artefacts used as input to the study are described. The process of the experimental study is described in the second subsection, consisting of pre-training followed by three activities: problem extraction done by individuals, problem filtering performed by individuals and problem merging performed by pairs. The last subsection describes how we analysed the data.

4.1 Study material: Observational reports and video clips from a usability evaluation

Prior to this study, an e-learning platform had been evaluated for usability with representative end-users. In that study, participants were users who carried out two tasks: (i) Provide and offer Learning Resources (see Figure 2) and (ii) Browse the catalogue options. For the former task a user was asked to register a learning resource, such as slides, video, text or other education material or education activity. For the latter task, a user was asked to browse the catalogue of learning resources. Browsing could be done according to several constraints, such as authors, language, category of topics, etc.

Figure 2 Educanext e-learning platform – Provide a new Learning Resource

Two different types of data collected in that study were used as input data to this study: video clips of onscreen activities with users' verbal comments and observational reports written by a usability specialist who made detailed notes of users' behaviour. The observational reports given to the participants of this study, playing the role of evaluators, were from two users, each carrying out the two different tasks of the e-learning platform. An excerpt of an observation report is shown in Figure 3.

Figure 3 An excerpt from an observational report from a usability evaluation – Provide and Offer Educational Material

4.2 Participants

Eight students of software engineering and computer science participated, seven male and one female. All but one were undergraduates and one was a graduate. Self-assessed knowledge in Human Computer Interaction (HCI) and experience of usability evaluations was rated on a scale from 1 to 5 (very poor to very rich). To find about participants' familiarity with the application domain, we asked them how much knowledge they had of e-Learning systems. Four of the participants had medium knowledge, two had poor knowledge, and one each had very poor and rich knowledge.

The aim of this study was to gain insight into participants' justification of their decisions to extract, filter and merge problems. As in the case of qualitative studies, it is not the aim of this study to generalize broadly. Although it is generally accepted that qualitative studies neither require random samples of participants nor as many participants as quantitative studies (Hennink, Hutter, & Bailey, 2010), the number of participants for usability evaluation is an ongoing issue (Marshall, Cardon, Poddar, & Fontenot, 2013). Because of the rather uniform set of participants and because of the extent of the analysis required, we found it justifiable to involve eight participants, who were interviewed after each of the three activities in the study, totalling 24 interviews. In each of these interviews, participants were asked six questions on two to three UPs. With the aim of creating guidelines for sample size of qualitative research, Marshall et al. (2013) conducted a study to analyse the number of interviews and interviewees and to analyse authors' justifications for the numbers. Their conclusion was that researchers seldom justified the sample size of their studies and there was a large variation in them. Their finding stated that while fewer than 20 interviews would risk not reaching saturation more than 40 interviews could lead to a loss of researchers' attention to analysis and reporting. We argue that such reduced attention would compromise the quality of these research activities. Overall, Marshall et al (2013) concluded that grounded theory qualitative studies should generally include between 20 and 30 interviews.

4.3 Procedure

Participants worked both individually and then collaboratively in four pairs where participants were randomly assigned to one of the pairs. The first phase of the study (Table 4) started with a pre-test training to familiarize participants with the platform and to strengthen their knowledge about the think-aloud user technique. The main phase of the study was conducted in the three activities described in the following subsections.

Table 4 Four activities participants participated in

4.3.1 Problem extraction

In the problem extraction process, participants were given four narrative observational reports (printed texts) of comparable length (see an example in Figure 3). Participants were individually required to analyse the four sets of reports and extract as many UPs as possible and register in a structured form including UP Description, Criteria applied, Severity level and Confidence level. To aid the participants in extracting UPs, we gave them the criteria for UP that John and Mashyna (1997) used for extracting UPs from tapes of evaluation sessions. These criteria were used in several studies, including (Niels Ebbe Jacobsen, Morten Hertzum, & Bonnie E John, 1998a; Law, 2006). Severity is a de facto standard in UP extraction and confidence is to gauge how sure usability evaluators are of their decisions (Hertzum, Jacobsen, & Molich, 2002). The constructs of the form are described below:

- *UP Description*: Narrative text based on the data from the observational reports.
- *Criteria applied:* Identify criteria justifying a UP according to a predefined scale (John & Mashyna, 1997), or name their own criteria to which the UPs could pertain. The criteria were the following:

  C1) Cannot continue without external help;

  C2) Tries several things and then explicitly gives up;

C3) Fails to achieve it or gets a wrong output;

C4) Commits an error that makes him/her pause for thought before he/she can continue (i.e. the duration of the pause is an indicator of problem severity);

C5) Expresses frustration, anger or surprise;

C6) Makes some negative comments on an interface element or proposes a design alternative.

- *Severity level:* Rate a UP severity as severe, moderate or minor, according to a predefined scale (Artim, 2003):

Severe usability problems are those that prevent the user from completing a task or result in catastrophic loss of data or time. Catastrophic loss of data implies either that the lost data cannot be reconstructed or that there is a very high cost to reconstruction. Catastrophic loss of time must be considered in light of the task duration.

Moderate usability problems are those that significantly hinder task completion but for which the user can find a work-around.

Minor usability problems are those that are irritating to the user but do not significantly hinder task completion.

- *Confidence level:* Indicate a participant's confidence that his or her identification of a particular UP was valid, on a 5-point Likert scale (1 = very low, 5 = very high).

After completing the problem extraction, each participant was interviewed to discuss the characteristics of three UPs randomly picked from his/her whole set of UPs. The interviews were semi-structured with six questions (see Table 5 in the Appendix). To gauge the participant's argumentation for the problem extraction, we asked why he/she considered this a UP. We wanted to know what resources he/she used when making the decision and thus we asked if the participant had experienced the same UP him-/herself or if he/she had seen it in the video. Sometimes the participants described a UP as a redesign proposal and thus we asked if the participant had any idea on how to improve the UP. To learn about the participants' arguments for their severity assessment and confidence levels, we asked them to give reasons for those ratings.

4.3.2 Problem consolidation by individuals

As part of the consolidation of UPs by individuals, each participant was asked to filter out any duplicate within each of the two tasks and merge similar UPs. The input to this activity was the outcome of the previous step, problem extraction. The participants identified UPs as retained, merged or discarded during this process. The consolidation strategies were collected from participants through semi-structured interviews that included three questions (see Table 6 in the Appendix). We asked how the participant had decided whether two UP descriptions were similar or different and how easy or difficult it had been to come to that conclusion.

4.3.3 Problem consolidation collaboratively by pairs

In the activity of consolidating UPs collaboratively by pairs, participants were required to consolidate their respective lists of UPs prepared in the individual sessions and make a master list of the UPs identified. The collaborative problem consolidation took place several days after the individual filtering step. The participants were randomly assigned to four pairs to merge their respective lists from the individual sessions into a master list. All materials used in the earlier sessions were accessible. Every item (i.e. single UP or merged UPs) in their own consolidated list was recorded in a structured form indicating which of the three possible changes was made: retained, discarded or merged. Following the consolidation exercise the groups were interviewed and asked three questions to reflect on the process (see Table 7 in the Appendix). Participants were asked to report strategies for deciding if two UPs were similar or different and how they decided on the severity ratings. We also asked how easy or difficult it was for the participants to convince their partner about the decision of a UP. No time limit was imposed on any of the procedures.

4.4 Data Analysis

To understand the main categories and concepts in participants' answers and in line with the explorative aim of the study, we analysed the data qualitatively using content analysis. We used conventional content analysis that starts with observation, after which codes are defined during data analysis and, lastly, codes are derived from data (Hsieh & Shannon, 2005; Krippendorff, 2012). In this study, the analytical process entailed breaking down the text from participants' answers in the semi-structured interviews into raw data (quotes) and then coding the raw data. The codes were reviewed and codes of similar statements combined as much as possible without losing the meaning of the content. Furthermore, the results were analysed and networks were used to formulate concepts into units of meaning and make categories from selective coding statements. Descriptive concepts were re-evaluated for interrelationships and gradually classified into a higher order of categories. Networks were formed and displayed visually to encourage communication in the research team. To ensure rigour, strategies such as audit trail, member checking, reflexivity and negative case analysis were adopted (Maschi, 2016).

The software ATLAS.ti, Qualitative Data Analysis (version: WIN 6.2) ("Atlas.ti ", 2016) was used for coding, classifying and organizing the data. ATLAS.ti is a tool for supporting the process of qualitative data analysis belonging to the family of CAQDAS programs (Computer-Aided Qualitative Data Analysis Software). The process of coding was to select text (quotes) from a document; code by short descriptions; classify codes and organize as structured information (Friese, 2012). Figure 4 shows an example of the coding of the question "Why did you consider this a UP?" There were four codes: "Scarce Feedback from the system", "Perception-Mental Model/Situation Awareness" and "Action – User knows what to do but can't in the UI". Example quotes behind the code "Perception – Mental model/Situation Awareness" were "User did something he thought was right, but he did not need to do this" or "lack of knowledge, user did not know what booking and access meant and LR concepts".

For the argument modelling, usability problems and individual answers in the semi-structured interviews were coded and translated into Toulmin theory of argumentation (see Figure 1). These models were then abstracted into higher level argumentation templates.

5 Results

5.1 What justifications and support do evaluators give for UP extraction, and how do they qualify their decisions?

In this section, we present the data required for answering the first research question (RQ1): What justifications evaluators give for extracting UPs and how they qualify their decisions. Eight participants extracted 71 problems from the two observational reports. Half of the problems were severe (54%), a quarter moderate (26%) and one fifth was rated minor (20%). Each participant was asked to pick three UPs of the lowest, medium and highest confidence level, and to answer questions on these. Categories were made of answers and arguments given by participants on why they considered a particular problematic instance to be a UP, i.e. why they extracted the problem, reasons for the severity and confidence ratings, how they felt about the video clips and if they had any suggestions for redesigns that would remove a UP.

5.1.1 Reasons for UP extraction - reasoning patterns

The first question (see Table 5 in the Appendix), *Why did you consider this a UP?* was asked to examine reasons for considering a user's problem in interacting with the system as a UP. The answers were analysed and grouped into a network of categories during a qualitative analysis (see Figure 4).

Figure 4 Why did you consider this a UP? - Network of answers (partially collapsed view)

In Figure 4 the root of the network or tree in this case is the question put forward, i.e. “Why did you consider this a UP?” Below the root come four categories in as many nodes in the network that emerged when the answers were coded. Three main categories emerged with the frequency in brackets: Scarce feedback from the system (5); Action (a user knows what to do but cannot do it in the UI) (8); Users’ perception that did not match their mental model or lack of situation awareness (9). Other answers did not form a distinct category (2). For the sake of brevity, three of the categories (nodes) are collapsed in Figure 4, but the category Perception – Mental Model / Situation Awareness is expanded. In that category we see all the answers to the question “Why did you consider this a UP?” e.g. “Very hindering. User didn’t know what was going on”.

The category of “Scarce feedback from the system” consisted of a user commenting on not being happy with the feedback from the system. A search button was missing and a full name was required for searching, so users thought that typing a search string instead of entering one letter at one time was time consuming. In one instance, an author was suggested to be a default author, and in another there was a lack of feedback on errors and insufficient information. Finally, the current state in the UI was not always clear for the user, like a problematic placement of a "Click here" link did not correspond to mutual user and participant experiences.

The category of “Action”, where a user knew what to do but could not, showed that a user could not finish his task by choosing categories, did not find Booking and did not notice the "Click here" link. Difficulties performing a task were mutual user and participant experiences and they thought Advanced Search was irritating and time-consuming.

The category “Perception – Mental Model / Situation awareness” revealed that a user had problems with the concept of a contributor and understanding messages. A user did something he thought was right, but did not notice what he was doing by searching in a subcategory instead of the whole database. A user got no results, did not realize how the system performed a search and was unfamiliar with the system filter. A user was unfamiliar with the Booking and Learning Resources concepts in the system. Furthermore, not knowing what was going on hindered the user, the system lacked feedback on errors, and there was a run time exception. In one instance, the participant did not articulate further on the reason for being a UP, but simply said that it should be fixed, but it was not a serious problem.

The answers to the question *“Why did you consider this a UP?”* were modelled using Toulmin model. An example model of a claim for making a UP is shown in Figure 5. The analysis of argumentation models derived from the answers showed that warrants were identifiable leading to clear claims. In all but one case the claim was: The UP is a usability problem. In one case the claim was: The UP is not a usability problem. This was a run time exception that could not be reproduced.

———————————————————————————————

Figure 5 An example argument model for UP extraction reasoning pattern

———————————————————————————————

In Toulmin model, justification of an argument is expressed as a warrant. A further analysis of the source of the warrant showed that, participants used users’ behaviour as justifications with phrases like *“The user”* or *“She”* or *“He”* (16 cases). In one case the participant referred explicitly to her own experience. In another case a participant used a video clip as justification for his answer. Worth mentioning is that in two instances a participant suggested a specific redesign, e.g. *“Needs more feedback, error messages”*. This too can be seen as a justification, i.e. the participant justifies her decision and emphasizes it with a redesign proposal. Although not as strong as statistics, guidelines or criteria, participants' references to users, own experience and redesign questions can be viewed as backings.

The criteria (C1-C6) participants registered as a part of their problem extraction were seen as backings to warrants, giving them additional support. In all cases when a problem was claimed as a usability problem, we could see that participants had identified a criterion during problem extraction. One might have expected that answers to the question: "*Why did you consider this a UP?"* would have included explicit references to one of the six criteria given to participants. However, participants never articulated such backing explicitly as answers to that question. To further explore this issue, we analysed the answers to see if participants had mentioned the criteria (C1-C6) implicitly. First, each of the first and third authors analysed the answers individually and then compared their results and came to a consensus. The results showed that around half of the answers were in agreement with the criteria stated explicitly by participants when asked, i.e. in 11 answers out of 23 UPs participants used implicit backing of the same criteria as they had stated for the UPs.

In two cases the participant gave qualifying statements of their argument, strengthening their claim. In one case saying *"Both me and the two users experienced…"* and in another one *"Both users had problems"*. A participant's confidence in extracting a UP can be seen as an explicit qualifier and is discussed separately in section 5.2.

The participants almost never gave any rebuttals to their claims. Three exceptions were when the participant thought that perhaps the user did not have the right background, the user did not understand the particular task, or the user did not have enough knowledge. Hence, in this third example the UP would not be expected to show up in the normal user group. A template of argument models for problem extraction is given in Table 8.

Table 8 Problem extraction - Argument model template

5.1.2 Redesign proposals

Participants were individually asked about ideas for UP improvements in the semi-structured interviews: *Do you have any idea how to improve this UP?* The analysis revealed four categories (number of instances): Help/Assistance/Automation (6); Navigation (6); Presentation (5) and Dialogue (6).

Suggestions in the group of Help/Assistance/Automation were e.g. about improved assistance and automation on how to use the search, to make a direct link to a help-window, to show where to search, to add the author automatically and to simplify the visualisation of the authors. In the navigation category participants suggested to add a search button, add a go button or a link to type a search string. Furthermore, there were suggestions to move a link to the word itself, make a tree-structure indicating a user’s position in the navigation structure of the UI and make a button instead of a link. Items in the category of Presentation were about visual improvements such as bigger fonts, better placement of a search button, more visual and clearer information, improved text and better separated and separate data coming in and out of the system. Finally, suggestions regarding Dialogue were for improved feedback if a search did not find anything, separating steps for booking and accessing, simplifying the search in Browse the Catalogue, providing clearer error messages to describe what happened and more information when there is only one contributor.

5.1.3 Prior experiences of a UP

Participants were asked if they had prior experiences of particular UPs by asking: *Have you experienced this kind of UPs when you worked with the system?* The results showed divided experiences, where in eleven cases participants acknowledged that they had experiences of a UP and in thirteen cases they reported no experiences.

5.1.4 Video-clips

When participants were asked what they thought about the video-clips, they gave positive feedback on using the video-clips for 14 UPs. One replied with a negative feedback and in nine cases no answer was provided. Interestingly, in two cases participants mentioned that listening to what users said in the video was useful and both listening and reading the text helped them. Furthermore, it helped participants to understand how users can have a very different background from their own. In one case, a participant remarked that he enjoyed watching the video-clips, which may indicate that a participant was more motivated in watching the videos than reading the text. Possibly this might ponder the question if using videos drew participant's attention from the UPs to user behaviour characteristics.

5.2 Evaluator's confidence

In this section we present the findings related to the second research question (RQ2): How evaluators explain how confident they were of the UPs identified. Translated into Toulmin theory, a confidence level is a qualifier of a claim acknowledging that a particular user's problem is a UP. Having extracted a problem and rated their confidence in the problem, participants were asked why they gave it a particular confidence rating, i.e. as very low to very high (on a five-point Likert scale) (see question 4, Table 5 in the Appendix). Participants' answers to this question were classified into six categories of reasoning patterns that were divided into two groups (see Table 9). In the first group participants assessed their confidence level according to their own understanding or experiences, or offered solutions to the problems. In the second group they assessed users' performance or expertise or used characteristics of the problems such as priority or severity to rate the confidence level.

Table 9 Reasoning patterns when rating UP confidence level

When explored further, the two groups showed different trends of ratings (see Figure 6). In the first one, where participants relied on their own experiences or intuitions, the ratings were more in the upper levels (10 out of 14 UPs on levels 3-5), with five UPs on level five. In the second group, where a participant assessed a user's experience or characteristics of problems, the ratings showed opposite trends in the lower confidence levels of one to three (9 out of 10 UPs). Thus, participants were more certain about the credibility of UPs when referring to their own experiences than when referring to users' performance or problem characteristics. An example of the former category, only including ratings on level 5 was: *"Think it is obvious that there should be a kind of tree-structure indicating user placement in the UI"*. On the other hand, when assessing a user or problem characteristics, it was more about distinguishing the roots of the UPs as here: *"Did not have this problem myself. Do not know whether to blame the user or the system"* (level 3).

Figure 6 UP confidence level according to reasoning patterns

It is worth pointing out how a participant's experience and users' expertise determine participants' confidence in a problem. Data on higher confidence level ratings (3-5) were characterized by participant experiences such as *"Because I had a similar problem"* and *"I saw that the user made this mistake and I made it myself"*. On the other hand, answers for lower confidence level ratings (1-2) referred to inexperienced users with examples such as *"User very confused and did not figure things out. I thought he misread something"*.

5.3 How do evaluators consolidate UPs in an individual vs. a collaborative setting and how easy did they think it is?

5.3.1 Consolidation

In this and the following sub-sections we attempt to answer the third research question (RQ3): How do evaluators consolidate UPs in an individual vs. a collaborative setting and how easy do they think it is. Participants filtered problems individually and merged them in pairs. During the individual filtering, 2 problems were marked as discarded, 12 as filtered and 57 retained. In the collaborative merging 6 problems were marked as discarded, 45 as merged, 17 as retained and three problems were missing from the merging. For the consolidation by individuals, 16 UP pairs were analysed and in the collaborative merging 8 UP pairs were analysed. Participants answered questions on why they decided that UPs were the same or not, how easy it had been and, additionally in the collaborative setting, why they rated the severity as they did (see Tables 6 and 7 in the Appendix).

Of the 16 problem pairs studied in the individual filtering, eight claims (two claims deduced from one instance set) resulted from the analysis of arguments but in nine cases no claims could be deduced. The data led to claims stating that two UPs were identical or not. Warrants of arguments for those claims included e.g. participants stating characteristics (tasks, severity, criteria) for UPs needing to be identical. In one argument, since the participant said that he or she thought that most users could finish this particular task, there was some uncertainty regarding the claim, demanding a qualifier - "*presumably*". This qualifier required a rebuttal: *"Unless the user cannot finish this task"*. In many of the nine cases where no claim could be deduced, participants explained how they filtered problems by citing a method, e.g. *"I looked at my descriptions and I read the text on both sides"*. On the other hand, Figure 7 shows an example of a pattern of an argument claiming that the two UPs are not the same. The data included a list of UPs extracted. When probed the participants replied that *"The UPs have no relations. Users are doing different things in different ways.* Then, a warrant implied in the argument for this claim was *"Two UPs need to be related to the same task to be identical"*.

____

Figure 7 A model of an argument in the individual filtering process

____

In the collaborative merging, six claims, from the eight problem pairs studied (two claims deduced from two instance sets), resulted from the analysis of answers to the question how participants decided to consolidate problems. In four problem pairs no claims could be extracted. All arguments resulted in claims stating that a situation is a usability problem.

In three of the cases a qualifier was added to an argument. In one case participants were unsure that a given description was a UP because it could be a problem in the operating system. In another case the participants were uncertain: *"we had difficulties deciding whether or not it was a UP – I became less certain"* which called for a qualifier, *("presumably"),* and a rebuttal *("unless the user is misunderstanding the task")*. In the third case a participant expressed certainty: *"It is definitely a UP, ..."* and we therefore added a qualifier stating this.

As an example of an argument on consolidation, two participants gave separate arguments to the question on why they had merged two problems. The difference in participants' views and confidence became apparent. One participant thought that *"this was the user's fault rather than a UP, but in fact fixable"* and the other one thought "*that this was not the system's fault but a problem in the operating system"* and that *"the user lacked knowledge and hence the UP became visible"*. Therefore, the argumentation model had a weakening qualifier (*"presumably")* and a rebuttal (*"there is a problem in the operating system"*). In another case where participants expressed difficulties in coming to a conclusion because of user misunderstanding, a participant qualified his decision with *"I became less certain"*. And explicitly said that he had not searched for further data: *"Did not want to watch the video – didn't even think about it"*.

The answers to the question how participants decided if two UP descriptions were similar or not, were categorised into reasoning patterns (see Table 10). In the first category, looking at tasks, context or otherwise referring to the user, participants referred to what a user was doing, what was hindering him/her in dealing with the tasks and in what context the problem occurred. The second category referred to participants evaluating characteristics of problems, including criteria for problem extraction, severity and/or a confidence level of the problem. The third category presented a method of reading text without referring to the content of that text, meaning that participants read the text that was given to them at the onset of the activity that contained descriptions of the usability evaluation or they read notes they had written themselves. Answers in the fifth category had no articulation of a decision and did not refer to any strategy.

____

Table 10 Reasoning patterns for whether two UPs were similar or not

____

In the individual problem consolidation, participants mainly looked at tasks, context or otherwise referred to a user and read text descriptions. The data showed that when looking at tasks or contexts participants gave the most intelligible answers on how they decided if two UPs were the same or not. The answers revealed what they saw and how they explored the problem. On the other hand, when reading text descriptions, the arguments were weak and did not include warrants justifying claims. In four cases participants were unable to articulate their decision.

The answers in the collaborative merging were categorised the same way as in the individual filtering (see Table 10). When looking at tasks, context or otherwise referring to a user in the collaborative effort, participants mainly argued whether or not a problem was a UP from describing their view of the users' behaviour. The second category, using characteristics of problems including criteria for problem extraction, severity and/or a confidence level of the problem, showed participants using severity ratings to come to a consensus. In some of those cases the participants agreed that an interaction constituted a problem but differed in the severity ratings. In three cases participants discussed whether a UP was a problem or not without explicitly articulating their decision method.

In the collaborative setting there is an indication to derive more claims (6 out of 8) over the individual filtering situation (8 out of 16), and we could see that the arguments tend to have more qualifiers (3) and rebuttals (2) in the collaborative merging than in the individual filtering where we have one qualifier and one rebuttal. It was more difficult to model the warrant in the argumentation of the pairs than in the individual filtering. In cases where the participants agreed, this was not so difficult, but in cases where the discussions were controversial, they explained each other's side, but did not say why they came to a consensus or what the justification for the UP was.

5.3.2 Ease of decision to consolidate problems and convince partner

Participants answered a question about how easy/difficult it was to make decisions regarding usability problem consolidation. The results showed that during individual filtering participants thought it was easy to make decisions regarding usability problem consolidation. In six cases of 16, they said that it was very easy to make decisions, and in another six they said it was easy. In only two cases participants said that it was difficult to make decisions.

After collaborative merging, participants were asked how easy or difficult it was for them to convince their partner about a judgment of a UP duplicate. Conversely, participants were asked how easy/difficult it was for them to be convinced by a partner about their judgment of a UP and they were asked to illustrate with some examples. The answers to these questions were mostly the same, showing it easy for them to convince their partner and to be convinced. In six cases of eight problem sets it was easy, and in the other two cases answers were not provided. An example answer was: "*It went quite well. We agreed on most things and we had very similar UPs. Two of them were exactly the same. The only thing that was different was UP.*"

5.4 How do evaluators explain the severity ratings of problems extracted and how do they analyse the severity of consolidated UPs in a collaborative setting?

5.4.1 Severity of a problem

In this and the next sub-section will answer the fourth research question (RQ4): How do evaluators explain the severity ratings of problems. In the UP extraction process, participants were asked: *Why did you rate the UP severity as minor, moderate or severe?* Figure 8 shows an example of an argument model for rating a UP as severe. In this example there is neither a qualifier nor a rebuttal. Participants were competent in rating the severity and, with two exceptions, capable of explaining their decisions. In one case it turned out that a participant changed his mind and decided that it was not a UP and in another one the user could not articulate an argument.

Figure 8 A model of an argument in rating severity of problems during individual extraction

In one of the 22 claims for severity, a participant gave a rebuttal for her decision. In this case, the participant thought that a problem might be unique to this user and the rebuttal called for a qualifier weakening the claim of severity. Such rebuttals are worth noting to avoid false UP ratings. In another case a participant was unsure whether the problem was moderate and hence the qualifier was used to weaken the claim. The participant said that he was unsure because he was inexperienced in using such a system and because of his misunderstanding of the severity ratings. A qualifier was used to strengthen the claim in one instance where a participant was particularly sure of her rating, saying that both she and the user had the same experience.

The warrants for the claims of severity ratings included statements about a user finishing a task and the time it took and statements saying that the user needed help from an instructor. In another answer participants referred to the mood of the user. Although references to parts of severity definitions could be seen in the arguments, they did not include explicit backings to severity definitions. A case that did not cite any of the severity definitions was when a participant related severity to the lack of a necessity for fixing a problem. Two of the claims were qualified, one with a weakening qualifier and one with a strengthening qualifier. Formulated in a rebuttal of an argument was a participant speculating that the problem was confined to this particular user. Summarising the argument models, a template for an argument for rating UP severity emerges and is shown in Table 11.

Table 11 Severity - Argument model template

The answers to the question of how participants decided the severity were categorised. We tried to use the same categories that emerged in the question: *How did you decide if two UPs were the same?* Almost all answers (19 / 24) fell into the category: Looking at tasks, context or otherwise referring to the user, and one answer belonged to the category: Characteristics of problem - Criteria for extraction or severity (see Table 12).

Table 12 Reasoning patterns on severity rating

5.4.2 Severity ratings in the collaborative settings

In this sub-section we present results on how participants analysed severity ratings of UPs in a collaborative setting and compare it to answers from a similar question asked during problem extraction (see Section 5.1). The data for the collaborative case comes from answers to question two in Table 7 (in the Appendix): *How did you decide on the severity rating of a UP?*

The collaborative results revealed participants using variable reasoning patterns on deciding the severity ratings. In three of eight cases, they looked at tasks, context or otherwise referred to the user, and in three other cases they discussed the ratings without answering how the ratings were decided. In the remaining two cases the dyads were unable to articulate their answer. In the individual problem extraction setting, participants almost exclusively used the method of looking at tasks, context or otherwise referring to the user to justify their severity. Thus, evaluating severity individually during problem extraction seemed clearer for participants than during the collaborative merging.

Of the answers on severity ratings in the collaborative setting only two answers out of eight led to claims on UP severity and those were for minor severity. In both of these incidents, participants looked at tasks or context or otherwise referred to the user.

6 Discussion and limitations

6.1 Discussion of findings

The results of this study showed that during usability problem extraction, novice usability evaluators can well justify their decisions on usability problem extraction, giving warrants leading to clear claims. At least implicitly as part of an argument, backings to warrants are seen in half the cases as a reference to predefined criteria of what is a usability problem. When asked to state criteria explicitly, novice usability evaluators could provide them. Criteria for defining a usability problem, implicitly mentioned in an argument or explicitly stated during problem extraction, can be viewed as a backing to a warrant. On the other hand, it is not entirely clear to what extent references to external resources such as the user, video clip or one's own experience as an expert provide backing to a warrant, because making the distinction between a warrant and its backing is not always easy and even some research has chosen not to distinguish between the two components of the argumentation (Stegmann et al., 2012). Novice usability evaluators rarely qualified their claims and seldom included rebuttals to claims. Qualifiers and rebuttals were seldom seen as a part of participants' arguments. This conclusion has to be viewed in the context that whereas participants had been prompted to select one of the six predefined criteria for their UPs, they had not been asked to give qualifiers and rebuttals. When asked explicitly about previous experience of a UP, less than half of the participants said that they had experienced a UP before. This finding agrees with previous results of studies (Nørgaard & Hornbæk, 2006; Yusop et al., 2016). Based on argumentation of participants, they seemed to be rather confident and only in rare cases did they qualify their claims to weaken or strengthen them. In accordance with a few weakening qualifiers, few rebuttals were given by novice usability evaluators, but when they were given, participants cited users' skills or background. Similar patterns have been found in practice (Nørgaard & Hornbæk, 2006). Boren and Ramey (2000) noted that practitioners used verbalization to gain a coarse level understanding of users' goals and motivations during the tasks and that they did not use it to help them build users' cognitive models. This coarse data may affect novice usability evaluators when arguing for the usability problem.

Novice usability evaluators backing arguments with their own experience were better convinced of their decisions than when grounding decisions by referring to users. Further research is needed to learn about the cause of this, e.g. if novice usability evaluators think that low usability is user’s fault (Norman & Nielsen, 2010) or if they need more training on how to use scientific data, i.e. results of usability evaluation, as evidenced in their backing (Murphy, Firetto, & Greene, 2017). Practitioners do not seem to look thoroughly at the data, since they rarely analysed verbalisations closely (Boren and Ramey, 2000).

The results showed that working individually or collaboratively required similar reasoning patterns in the consolidation process. In both the individual and collaborative settings, participants had difficulty in giving justifications for their decisions on consolidation. This is in accordance with Rittel (1987) who showed that deliberations on design terminated with judgements. He concluded that evaluators may make up their minds without being able to derive reasoning from deliberations. Comparing how rational individuals vs. pairs were during consolidation, we see that pairs were slightly better at deriving claims, qualifying their decision and putting forth rebuttals. Thus, we did not see the same decisive benefits of working in teams as Maciejovsky et al. (2013) did. In studies where participants need to learn complex rules of games and somehow divide tasks among them, the benefit of working in groups may be larger (Kugler et al., 2012) than in our study where participants work on the same thing.

Previous research has shown that having collaborators explicitly state their own findings before collaborating, compared to implicitly deriving them, will help them during argumentation in the collaborative setting (Papadopoulos et al., 2013). Contrary to what Papadopoulos et al. (2013) found, writing down their own findings before discussing them collaboratively did not seem to help novice usability evaluators in this study when the UPs were different. Stegmann et al. (2012) investigated whether the quality of argumentation of a partner in a dyad affected the depth of cognitive elaboration, but their results were inconclusive. They speculated whether the difference in partners’ (novice usability evaluators in our case) positions would have an effect on their ability to argue, hypothesising that agreeing partners might not engage in cognitive elaboration whereas disagreeing partners might stimulate more discussion. Although the results of this study gave some indication of how evaluators argued similar vs. dissimilar problems, a more systematic study is needed to investigate if evaluators argue differently in the two cases.

Regarding how easy or difficult it was to come to a consensus concerning usability problem consolidation, both individuals and pairs said that it was easy to very easy. Recall that participants were asked to answer this question on two UPs they found most controversial. It is hard to say what could be the reason for their ease of convincing one another. The reported ease of the collaborative consolidation may possibly be an artefact of the study setup. Since the UPs were derived from the same material, i.e. the written reports, the UPs may not have raised enough conflict. Our study gave participants written reports instead of videos as was done in Jacobsen et al. (1998a). The reason was to avoid the evaluator effect, i.e. that evaluators would extract different UPs from the same material. Using video for extraction could also be too demanding for novice usability evaluators.

Because of the few problems examined during collaborative merging it is difficult to generalize on argumentation of severity from the discussions. Only in three of eight cases were participants able to articulate their decisions on severity and two answers led to claims on UP severity. There may be several reasons for this. The questions may have been unsuitable for stimulating such claims. Research has consistently shown that computer-supported collaboration scripts can help partners increase the quality of argumentation (Stegmann et al., 2012). Therefore, to argue their positions more effectively, novice usability evaluators might be supported with scripts including templates of claims, warrants and backings. The second reason for not being able to derive claims might be that the chosen Toulmin argumentation model may not be the most suitable one for explaining different arguments of evaluators in a dyad or in a group. Since Toulmin argumentation theory is a description of the outcome of the discussion, i.e. the product, it is not able to depict the process, i.e. how collaborators actually interact during argumentation (Nielsen, 2013).

6.2 Limitations of the study

There are several limitations of this study. The number of participants was only eight and the number of pairs during collaborative merging was four. However, the number of items was higher since individuals addressed three problems each and during consolidation one pair of problems each. Because of the low number of participants we could not verify that the categories are fully saturated. This can motivate future work in this area.

Another limitation concerns the argumentation models. The data were not facts in the sense of being a description of truth but observations of users' behaviour while interacting within a system. However, in the eyes of the observer data could be seen as the truth and at least he or she used it to conclude on the usability of the system. Erduran (2007), as cited in (Nielsen, 2013), has pointed out this methodological difficulty of distinguishing between the different elements of Toulmin theory, e.g. the difference between data and the warrant. In our analysis of the data, we thoroughly discussed the meaning of the components and consistently applied rules for deducing them.

Since expert usability evaluators are likely to have better knowledge on usability problems, the results might have been different if we had recruited experts. However, it has been stated that experts often cannot articulate their knowledge because their knowledge is tacit (Chi, 2006). In a future study it would be worthwhile to learn if experts have skills in argumentation and explication of tacit knowledge gained from experience in arguing usability problems to other designers or software developers. As we mentioned in our discussions in the previous section, we provided novice usability evaluators with written protocols of observations instead of asking them to observe users. In a future study, it could be interesting to see if this has an impact on their arguments.

In this study, we did not attempt to formally evaluate the quality of the arguments. In studies of argumentation skills in the field of education, especially science education, the quality of the arguments has been a visible focus. In an attempt to evaluate changes in argumentation skills, researchers are concerned with the methodological aspects of the evaluation of the argumentation skills, e.g. before and after an intervention that is meant to improve skills. An example of an analytical framework has been suggested by Erduran, Simon, and Osborne (2004) who used the inclusion of rebuttals in argumentation to define five levels of quality, where level one is the lowest quality with simple claims and level five consists of extended arguments with more than one rebuttal. Such a framework could be useful for future research, aiming to study gain in argumentation skills of novice usability evaluators after receiving relevant training in using argumentation templates.

In addressing the above limitations, we have encouraged researchers to do further work to get closer to answering the overall question of this paper. This could be done by using the current highlights as hypotheses in future work and by exploring new areas of study as suggested above.

7 Conclusion

The overall aim of the research was to learn about the evidence and validity presented by novice usability evaluators by exploring argumentation patterns that they provide individually and in teams during different phases of usability evaluation. The results showed that novice usability evaluators can well justify their decisions on usability problem extraction, giving warrants leading to clear claims, when probed. Backings to warrants are only seen in half the cases as a reference to predefined criteria of what is a usability problem. On the other hand, when asked to state criteria explicitly, thus giving a backing, evaluators could provide them in all cases. Novice usability evaluators rarely qualified their claims and seldom included rebuttals to claims. The results of this study showed that in both the individual and the collaborative settings, novice usability evaluators had difficulty in giving justifications for their decisions on consolidation. While novice usability evaluators could justify their rating of severity during UP extraction with occasional backing, it was difficult to generalize on argumentation of severity from the discussion on severity during collaborative merging because of the few problems examined. The experiments could elicit novice usability evaluators' reasoning patterns during problem extraction and consolidation as well as during their assessment of severity and confidence.

The results have provided three main contributions to research and practice. Forming reasoning patterns and formulating argument models have provided improved insight into how novice usability evaluators, individually and collaboratively, consolidate UPs. We know what kind of backing they give and where they have reservations about their decisions and why. As a research tool, Toulmin model enabled us to learn about the evidence and validity presented by novice usability evaluators. The model was understandable but it took some training and required discussion among the researchers to map novice usability evaluators' response in the domain of usability evaluations to individual segments of the model. The argumentation templates proposed in this paper could be valuable for future research in analysing dialogical argumentation during UP extraction and consolidation, but, as has been pointed out by Nielsen (2013), such templates would require additional complementing frameworks to describe the dialectical process.

The practical implications of the findings are to use argumentation templates as scripts to support individuals and groups while extracting, filtering and merging. This could be particularly valuable for novice usability evaluators. This result harmonises well with that of Nörgaard and Höegh (2008) who, after studying different forms of usability evaluation outcomes, recommended that usability evaluators back up the warrants behind usability claims. The argumentation templates could also fulfil a need found by Yusop et al. (2016), who concluded, after conducting a survey with software developers, that developers want usability problem descriptions to include the cause of a problem, a characteristic that is lacking, according to reporters of defects.

The findings of the study show that novice usability evaluators lack skills that could be improved with training. While novice usability evaluators seem to be able to extract problems and consolidate them and given warrants for them, they seem to rely on own experience rather than citing user experiences when providing a backing and be less skilled in giving a backing, qualifying their claims and expressing rebuttals. Similar results were found in Dalsgaard et al. (2013) who studied design students. These are aspects that should be emphasised in novices' training, giving them better skills in communicating with other designers and customers, hopefully resulting in improved downstream utility (Law, 2006).

The argumentation models could be a first step in using argumentation-based decision support systems (Introne & Iandoli, 2014; Longo & Hederman, 2013; Noroozi et al., 2012; Rahwan, Zablith, & Reed, 2007) for usability evaluators. A prerequisite to this is further work to assess the quality of argumentation models for use in a decision support system for shared decision making.

8 References

Aristotle. (2006). *On Rhetoric: A theory of Civic Discourse* (G. A. Kennedy, Trans. 2nd ed.). Oxford: Oxford University Press.

Artim, J. (2003). Usability problem severity ratings. *Retrieved 2008.* Retrieved from http://www.primaryview.org/CommonDefinitions/Severity.html

Atlas.ti (2016). Retrieved from http://atlasti.com/

Blandford, A., Keith, S., & Fields, B. (2006). Claims Analysis "In the Wild:" A Case Study on Digital Library Development. *International Journal of Human–Computer Interaction, 21*(2), 197-218. doi:10.1207/s15327590ijhc2102_5

Boland, R. J. (2008). Decision Making and Sensemaking. In *Handbook on Decision Support Systems 1: Basic Themes* (pp. 55-63). Berlin, Heidelberg: Springer Berlin Heidelberg.

Boren, T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261-278. doi:10.1109/47.867942

Bornoe, N., & Stage, J. (2017). Active Involvement of Software Developers in Usability Engineering: Two Small-Scale Case Studies. In R. Bernhaupt, G. Dalvi, A. Joshi, D. K. Balkrishan, J. O'Neill, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part IV* (pp. 159-168). Cham: Springer International Publishing.

Carroll, J. M., & Rosson, M. B. (1992). Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems (TOIS), 10*(2), 181-212.

Chi, M. T. (2006). Two approaches to the study of experts' characteristics. *The Cambridge handbook of expertise and expert performance*, 21-30.

Clemmensen, T. (2011). Templates for Cross-Cultural and Culturally Specific Usability Testing: Results From Field Studies and Ethnographic Interviewing in Three Countries. *International journal of human-computer interaction, 27*(7), 634-669. doi:10.1080/10447318.2011.555303

Cockton, G., & Lavery, D. (1999). *A framework for usability problem extraction.* Paper presented at the Human-Computer Interaction - INTERACT'99, Edinburgh, UK.

Cross, N. (2001). Designerly Ways of Knowing: Design Discipline Versus Design Science. *Design Issues, 17*(3), 49-55. doi:10.1162/074793601750357196

Dalsgaard, P., Dindler, C., & Fritsch, J. (2013). Design argumentation in academic design education. *Nordes 2013: Experiments in design research, 5*.

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*: Intellect books.

Erduran, S. (2007). Methodological Foundations in Study of Argumentation in Science Education. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in Science Education-Perspectives from Classroom Based Research* (pp. 47-68): Springer Science.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science education, 88*(6), 915-933.

Fischer, G., Lemke, A. C., McCall, R., & Morch, A. I. (1991). Making argumentation serve design. *Human–Computer Interaction, 6*(3-4), 393-419.

Friese, S. (2012). Qualitative Data Analysis with ATLAS.ti.

Furniss, D. (2008). *Beyond problem identification: valuing methods in a 'system usability practice'.* UCL (University College London),

Følstad, A., Law, E., & Hornbæk, K. (2012). *Analysis in practical usability evaluation: A survey study.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Hennink, M., Hutter, I., & Bailey, A. (2010). *Qualitative research methods*: Sage.

Hertzum, M. (2006). Problem Prioritization in Usability Evaluation: From Severity Assessments Toward Impact on Design. *International journal of human-computer interaction, 21*(2), 125-146. doi:10.1207/s15327590ijhc2102_2

Hertzum, M., Borlund, P., & Kristoffersen, K. B. (2015). What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International journal of human-computer interaction, 31*(9), 557-570.

Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). *Usability inspections by groups of specialists: perceived agreement in spite of disparate observations*. Paper presented at the CHI '02 Extended Abstracts on Human Factors in Computing Systems, Minneapolis, Minnesota, USA.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology, 29*(1), 97-111.

Hornbæk, K., & Frøkjær, E. (2005). *Comparing usability problems and redesign proposals as input to practical systems development.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers, 20*(6), 505-514.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research, 15*(9), 1277-1288.

Introne, J., & Iandoli, L. (2014). Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence. *Decision Support Systems, 64*, 79-89.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR), 33*(4), 470-516.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998a). *The evaluator effect in usability studies: Problem detection and severity judgments.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998b). *The evaluator effect in usability tests*. Paper presented at the CHI 98 Conference Summary on Human Factors in Computing Systems, Los Angeles, California, USA.

John, B., & Mashyna, M. M. (1997). Evaluating a Multimedia Authoring Tool. *Journal of the American Society for Information Science (1986-1998), 48*(11), 1004-1023.

Johnson, P. M., & Tjahjono, D. (1998). Does every inspection really need a meeting? *Empirical Software Engineering, 3*(1), 9-35.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*: Sage.

Kugler, T., Kausel, E. E., & Kocher, M. G. (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(4), 471-482.

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review, 62*(2), 155-179.

Law, E. L.-C. (2006). Evaluating the downstream utility of user tests and examining the developer effect: A case study. *International journal of human-computer interaction, 21*(2), 147-172.

Law, E. L.-C., & Hvannberg, E. T. (2008). *Consolidating usability problems with novice evaluators.* Paper presented at the Proceedings of the 5th Nordic conference on Human-Computer Interaction: Building Bridges.

Longo, L., & Hederman, L. (2013). Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning. In K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, & N. Zhong (Eds.), *Brain and Health Informatics: International Conference, BHI 2013, Maebashi, Japan, October 29-31, 2013. Proceedings* (pp. 168-180). Cham: Springer International Publishing.

Maciejovsky, B., Sutter, M., Budescu, D. V., & Bernau, P. (2013). Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Management Science, 59*(6), 1255-1270.

MacLean, A., Young, R. M., Bellotti, V. M., & Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Human–Computer Interaction, 6*(3-4), 201-250.

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *Journal of Computer Information Systems, 54*(1), 11-22.

Maschi, T. (2016). Holistic Analysis, Discerning Meaning from Narrative and Numeric Data. In *Applying a Human Rights Approach to Social Work Research and Evaluation* (pp. 69-81): Springer.

Mercier, H. (2016). The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences, 20*(9), 689-700. doi:http://dx.doi.org/10.1016/j.tics.2016.07.001

Moran, T. P., & Carroll, J. M. (1996). Overview of design rationale. *Design rationale: Concepts, techniques, and use*, 1-19.

Murphy, P. K., Firetto, C. M., & Greene, J. A. (2017). Enriching Students' Scientific Thinking Through Relational Reasoning: Seeking Evidence in Texts, Tasks, and Talk. *Educational Psychology Review, 29*(1), 105-117. doi:10.1007/s10648-016-9387-x

Nielsen, J. A. (2013). Dialectical features of students' argumentation: A critical review of argumentation studies in science education. *Research in Science Education, 43*(1), 371-393. doi:10.1007/s11165-011-9266-x

Norman, D. A., & Nielsen, J. (2010). Gestural interfaces: a step backward in usability. *interactions, 17*(5), 46-49.

Noroozi, O., Weinberger, A., Biemans, H. J., Mulder, M., & Chizari, M. (2012). Argumentation-based computer supported collaborative learning (ABCSCL): A synthesis of 15 years of research. *Educational Research Review, 7*(2), 79-106.

Nørgaard, M., & Hornbæk, K. (2006). *What do usability evaluators do in practice?: An explorative study of think-aloud testing.* Paper presented at the Proceedings of the 6th conference on Designing Interactive systems.

Nörgaard, M., & Höegh, R. T. (2008). *Evaluating usability: Using models of argumentation to improve persuasiveness of usability feedback*. Paper presented at the Proceedings of the 7th ACM conference on Designing interactive systems, Cape Town, South Africa.

Olson, G. M., Olson, J. S., Storrøsten, M., Carter, M., Herbsleb, J., & Rueter, H. (1996). *The structure of activity during design meetings*: L. Erlbaum Associates Inc.

Papadopoulos, P. M., Demetriadis, S. N., & Weinberger, A. (2013). 'Make it explicit!': Improving collaboration through increase of script coercion. *Journal of Computer Assisted Learning, 29*(4), 383-398. doi:10.1111/jcal.12014

Rahwan, I., Zablith, F., & Reed, C. (2007). Laying the foundations for a world wide argument web. *Artificial intelligence, 171*(10), 897-921.

Reed, C., Walton, D., & Macagno, F. (2007). Argument diagramming in logic, law and artificial intelligence. *The Knowledge Engineering Review, 22*(01), 87-109.

Rittel, H. W. (1987). *The reasoning of designers*. Paper presented at the Congress on Planning and Design Theory and

Schriftenreihe des Instituts fuer Grundlagen der Palnung, Universitaet Stuttgart 1988, Boston.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences, 4*(2), 155-169.

Sauer, C., Jeffery, D. R., Land, L., & Yetton, P. (2000). The effectiveness of software development technical reviews: A behaviorally motivated program of research. *Ieee Transactions on Software Engineering, 26*(1), 1-14. doi:Doi 10.1109/32.825763

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning, 5*(1), 43-102.

Shum, S. B. (1996a). Analyzing the usability of a design rationale notation. *Design rationale: Concepts, techniques, and use*, 185-215.

Shum, S. B. (1996b). Design argumentation as design rationale. *The encyclopedia of computer science and technology, 35*(20), 95-128.

Stegmann, K., Wecker, C., Weinberger, A., & Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science, 40*(2), 297-323. doi:10.1007/s11251-011-9174-5

Sutcliffe, A. G., & Carroll, J. M. (1999). Designing claims for reuse in interactive systems design. *International Journal of Human-Computer Studies, 50*(3), 213-241.

Toulmin, S. E. (1982). The construal of reality: Criticism in modern and postmodern science. *Critical Inquiry*, 93-111.

Toulmin, S. E. (2003). *The Uses of Argument*: Cambridge University Press.

Yusop, N. S. M., Grundy, J., & Vasa, R. (2016). *Reporting Usability Defects–Do Reporters Report What Software Developers Need?* Paper presented at the Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering.

Yusop, N. S. M., Grundy, J., & Vasa, R. (2017). Reporting Usability Defects: A Systematic Literature Review. *Ieee Transactions on Software Engineering, 43*(9), 848-867.

Appendix

Table 5 Questions after problem extraction by individuals

Table 6 Questions after consolidation by individuals

Table 7 Questions after consolidation by pairs

Figures

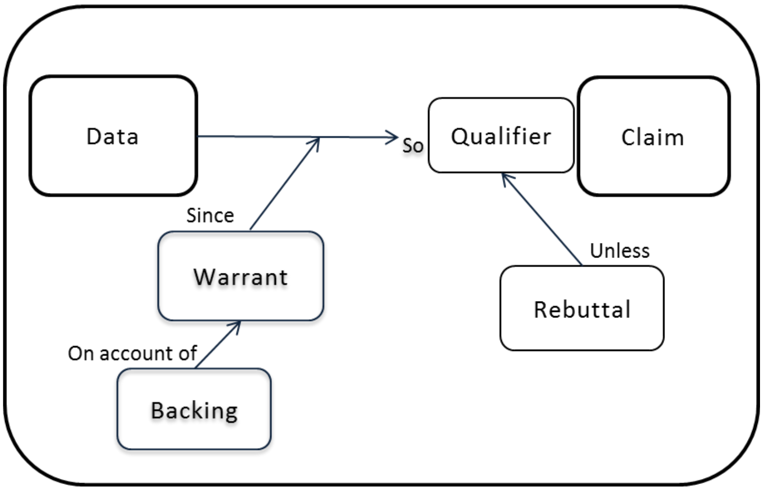
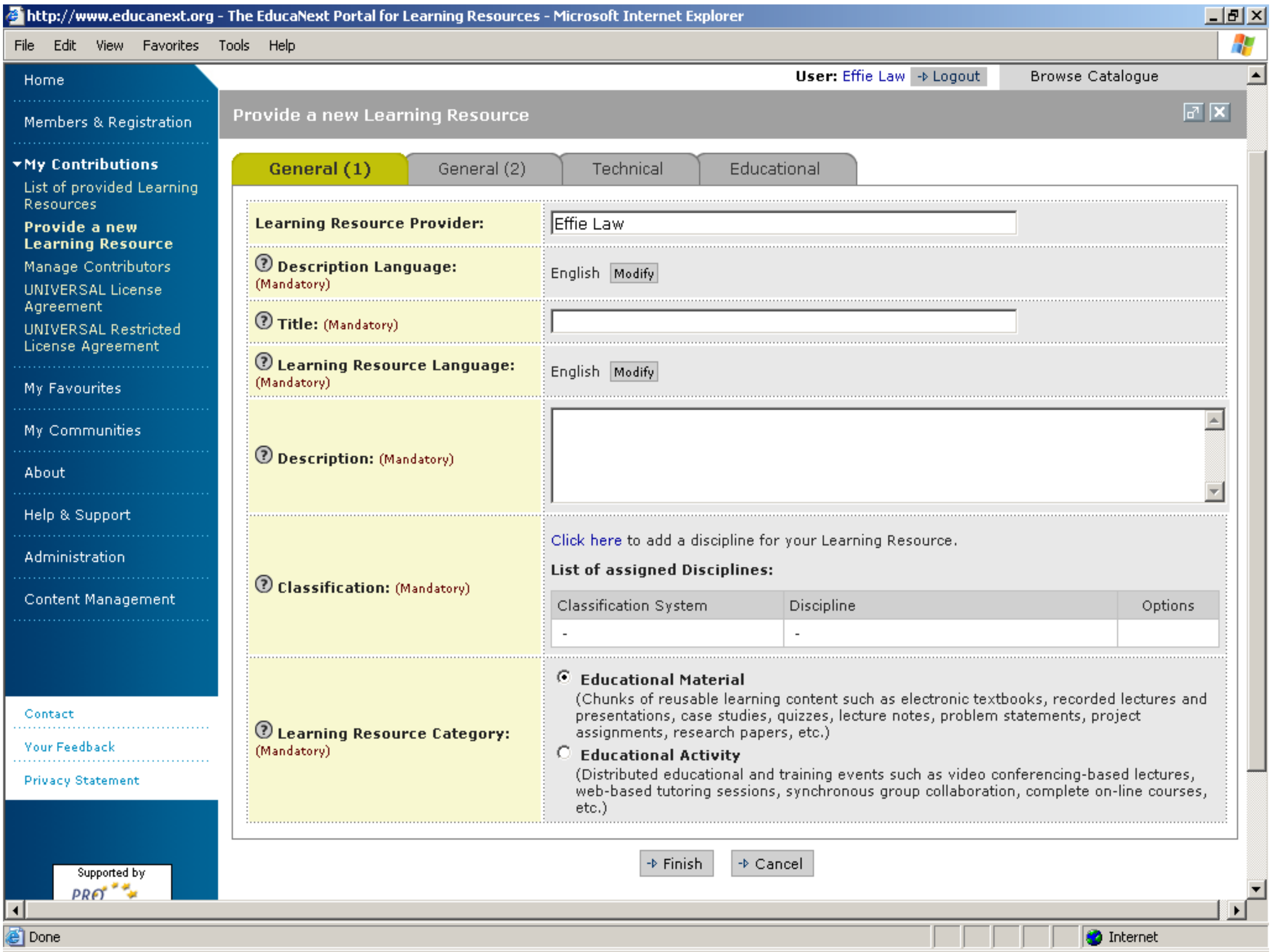


Figure 1 Toulmin's Theory: A model of an argument



Provide a new Learning Resource

Learning Resource Provider:	Effie Law
Description Language: (Mandatory)	English
Title: (Mandatory)	
Learning Resource Language: (Mandatory)	English
Description: (Mandatory)	
Classification: (Mandatory)	Click here to add a discipline for your Learning Resource. List of assigned Disciplines:
Learning Resource Category: (Mandatory)	● Educational Material (Chunks of reusable learning content such as electronic textbooks, recorded lectures and presentations, case studies, quizzes, lecture notes, problem statements, project assignments, research papers, etc.) ○ Educational Activity (Distributed educational and training events such as video conferencing-based lectures, web-based tutoring sessions, synchronous group collaboration, complete on-line courses, etc.)

Classification System	Discipline	Options
-	-	

Figure 2 Educanext e-learning platform – Provide a new Learning Resource

Task 2: Provide and Offer Educational Material

Completion:

Full.

Context: My Contribution

General (1): The user was a slightly confused about selecting a language twice i.e. Description language and LR language.

General (1): Classification field. The user didn't notice the click here link and attempted to click the form and asked the local tester whether this field included something (data) underlying. The local tester replied that this field could not be clicked directly. Then the user still tried to click it once more and then went to the next form i.e. General (2).

General (2). The user first searched for an author and typed in "s" in the first name field and then clicked the go back button by mistake and realised right away.

Then when the user was back in the form Search for or Create a new Contributor then she tried to click the bullet point in front of the Search for an existing contributor

Figure 3 An excerpt from an observational report from a usability evaluation – Provide and Offer Educational Material

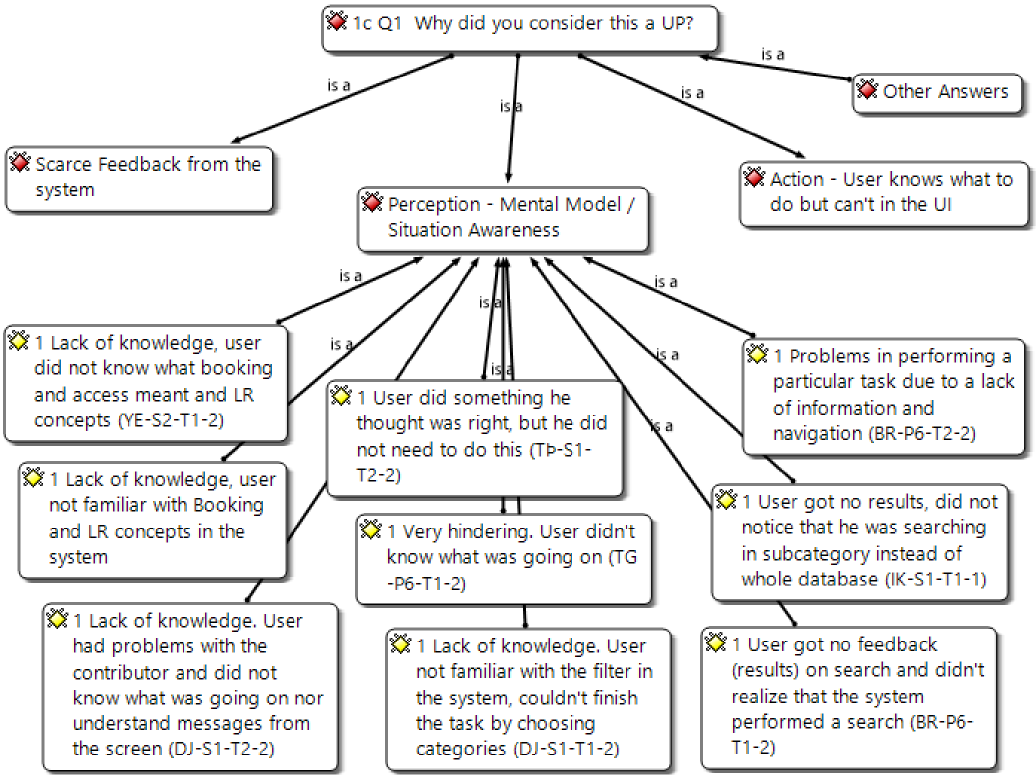

1c Q1 Why did you consider this a UP?
is a
Other Answers
Scarce Feedback from the system
Perception - Mental Model / Situation Awareness
Action - User knows what to do but can't in the UI
1 Lack of knowledge, user did not know what booking and access meant and LR concepts (YE-S2-T1-2)
1 User did something he thought was right, but he did not need to do this (TP-S1-T2-2)
1 Problems in performing a particular task due to a lack of information and navigation (BR-P6-T2-2)
1 Lack of knowledge, user not familiar with Booking and LR concepts in the system
1 Very hindering. User didn't know what was going on (TG -P6-T1-2)
1 User got no results, did not notice that he was searching in subcategory instead of whole database (IK-S1-T1-1)
1 Lack of knowledge. User had problems with the contributor and did not know what was going on nor understand messages from the screen (DJ-S1-T2-2)
1 Lack of knowledge. User not familiar with the filter in the system, couldn't finish the task by choosing categories (DJ-S1-T1-2)
1 User got no feedback (results) on search and didn't realize that the system performed a search (BR-P6-T1-2)

Figure 4 Why did you consider this a UP? - Network of answers (partially collapsed view)

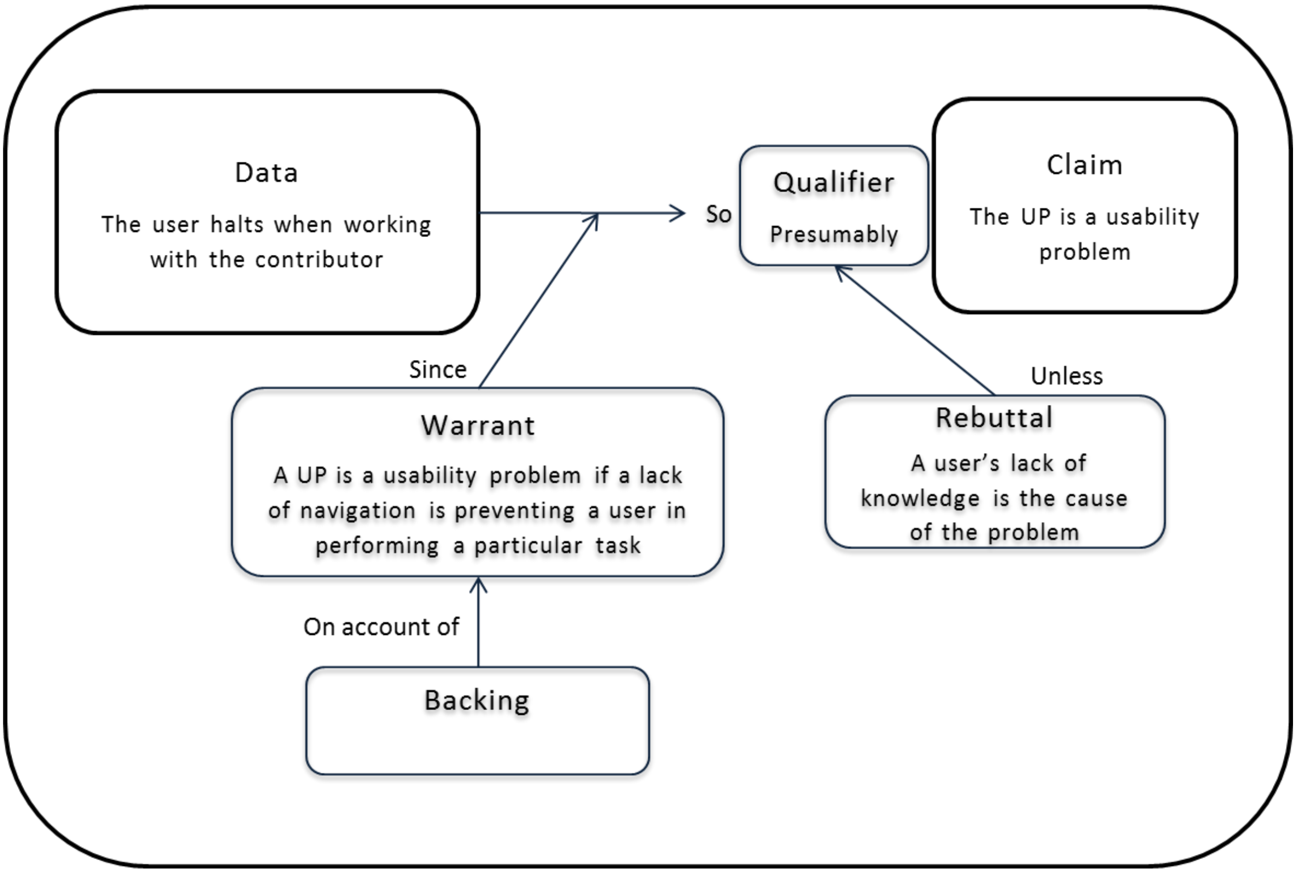

Data
The user halts when working with the contributor
So
Qualifier
Presumably
Claim
The UP is a usability problem
Since
Warrant
A UP is a usability problem if a lack of navigation is preventing a user in performing a particular task
Unless
Rebuttal
A user's lack of knowledge is the cause of the problem
On account of
Backing

Figure 5 An example argument model for UP extraction reasoning pattern

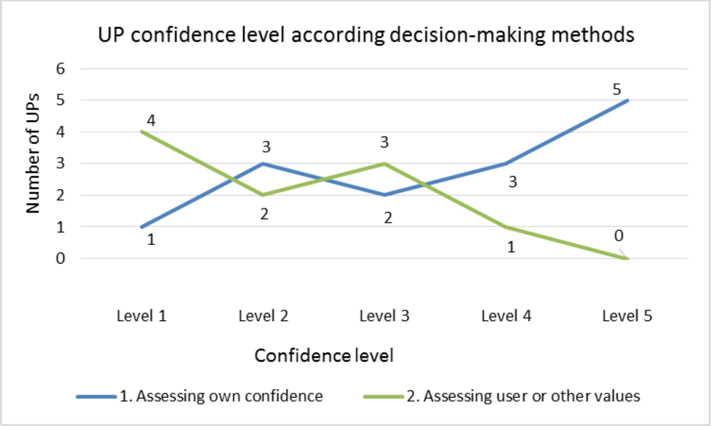


Figure 6 UP confidence level according to Reasoning patterns

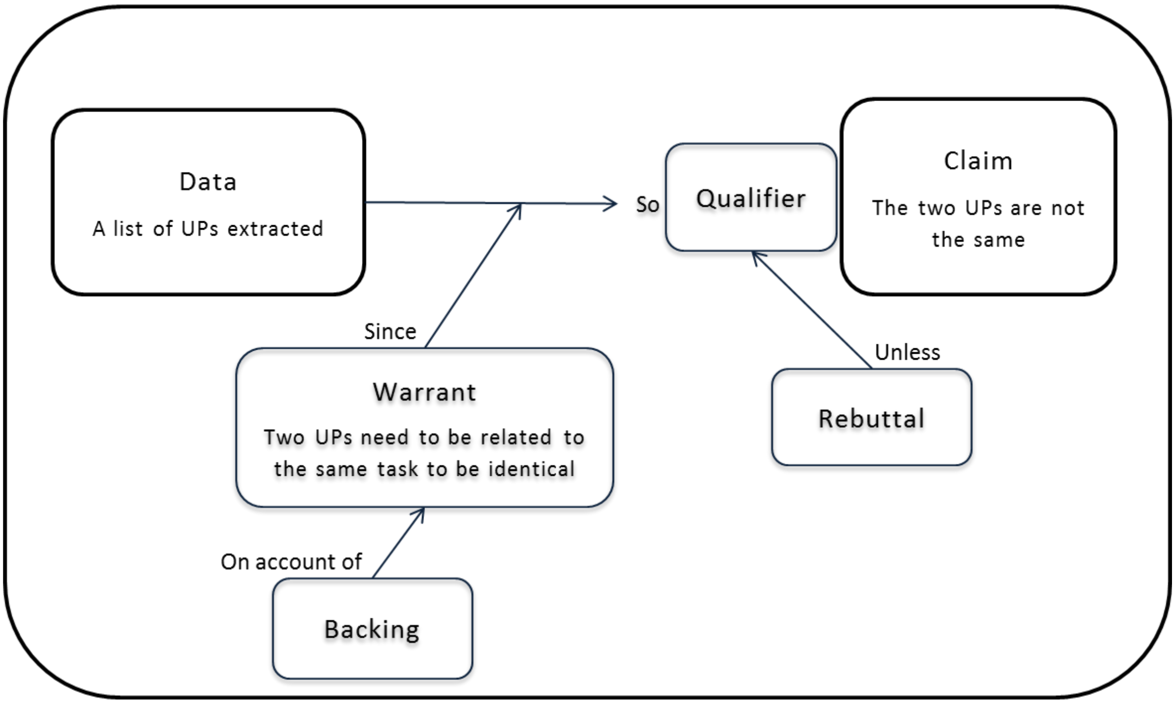


Figure 7 A model of an argument in the individual filtering process

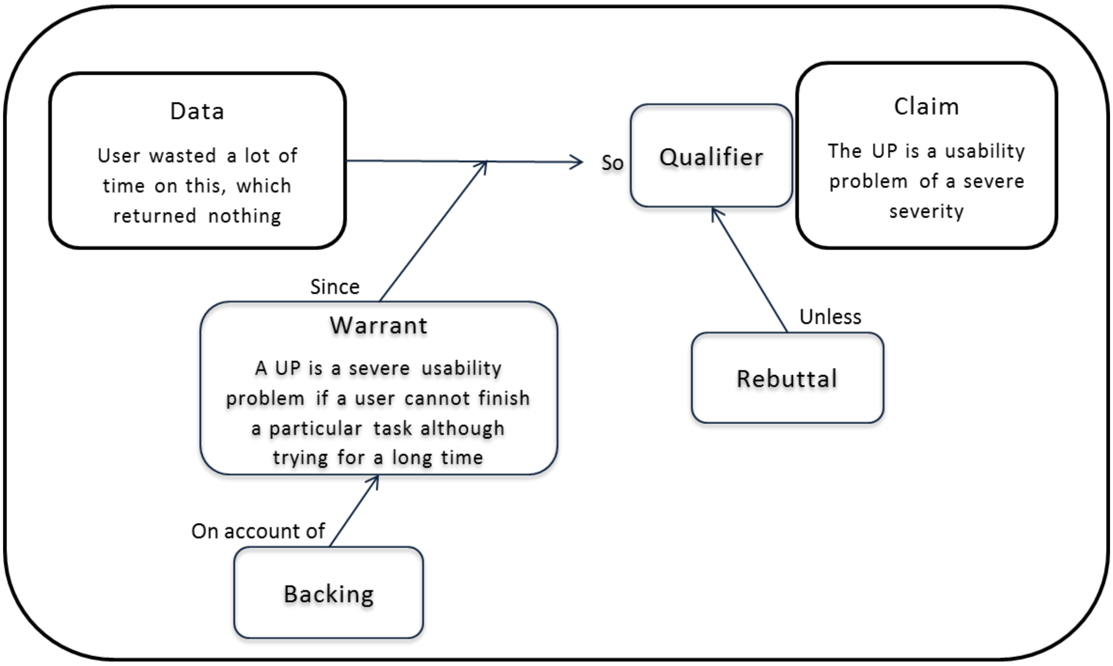

Data
User wasted a lot of time on this, which returned nothing
So
Qualifier
Claim
The UP is a usability problem of a severe severity
Since
Warrant
A UP is a severe usability problem if a user cannot finish a particular task although trying for a long time
Unless
Rebuttal
On account of
Backing

Figure 8 A model of an argument in rating severity of problems during individual extraction

Tables

Table 1 An example from data analysis leading to a model of an argument

Data	Warrant	Backing	Qualifier	Rebuttal	Claim
Missing search button	since a missing search button hinders a user in performing a particular task	None given	None given	None given	A UP description [missing search button] is a usability problem

Table 2 Characteristics collected and outcome of analysis during individual problem extraction

Data collected	Type of outcome of analysis	
Characteristic	Categories	Argument models
Problem extraction	Yes	Yes
Confidence	Yes	
Experienced before	Yes	
Severity	Yes	Yes
Redesign	Yes	

Table 3 Characteristics collected and outcome of analysis during individual filtering and collaborative merging

Data collected	**Type of outcome of analysis**	
Characteristic	**Categories**	**Argument models**
Consolidation	Yes	Yes
Ease of decision	Yes	
Severity – only for collaborative setting	Yes	Yes

Table 4 Four activities evaluators participated in

Activities
(1) Pre-test training phase
A lecture introducing the e-learning platform and user testing
Individually, participants read reference material about user testing
Individually, participants solve three tasks with the core functions of the platform
Collection of personal data (a pre-test questionnaire)
(2) Problem extraction, individual
Four sets of narrative observational reports analysed
Usability problems extracted according to predefined criteria
Usability problems assessed for severity and confidence and registered in a structured form
Semi-structured interviews of six questions discussing characteristics of three randomly picked Ups
(3) Consolidation, individual
Duplicates of similar UPs previously extracted within each task, filtered and merged
Unique UPs identified as retained, merged or discarded and reassessed
Semi-structured interviews of three questions discussing consolidation strategies
(4) Consolidation, collaborative performed by pairs
Consolidation of UPs from a master list of UPs prepared in the individual sessions
Master list registered and assessed in a structured form indicating changes of retained, merged or discarded UPs
Semi-structured interviews of three questions discussing consolidation strategies

Table 5 Questions after problem extraction by individuals

Problem Extraction - Individual

Questions for the three usability problems with the highest, average and lowest confidence level.

1. Why did you consider this a UP?
2. Have you experienced this kind of UP when you worked with the system?
3. Why did you rate the UP severity as (minor, moderate or severe)?
4. Why did you rate your confidence level as (very low … very high)?
5. What do you think about the video clips?

   Or, how useful did you find the video clips for discovering the UP?
6. Do you have any idea how to improve this UP?
7. Any follow-up question (specify):

Table 6 Questions after consolidation by individuals

	Problem Consolidation - Individual
1.	How many duplicates did you identify
2.	How did you decide whether the two UP descriptions were similar or different?
3.	How easy/difficult was it for you to make the decisions?

Table 7 Questions after consolidation by pairs

Problem Consolidation - Collaborative
1. How did you decide whether the two UP descriptions were similar or different? Please choose one UP that was most controversial.
2. How did you decide on the severity rating of a UP? Please choose one UP that was most controversial.
3. How easy/difficult was it for you to convince your partner about your judgment of a UP? Conversely, how easy/difficult was it for you to be convinced by your partner about her/his judgment of a UP? Please illustrate with some examples.
4. Any follow-up question (specify):

Table 8 Problem extraction - Argument model template

Data	Warrant	Backing	Qualifier	Rebuttal	Claim
A description of flaws in the interaction possibly having negative consequences for the user	A condition for a UP problem, e.g. a description of flaws in the interaction possibly having negative consequences for the user	One of six criteria (C1-C6) stated during problem extraction. Implicit criteria. References to users, own experiences, video-clip or redesign suggestions.	Two users or an evaluator and a user had the same experience, strengthening the claim. Presumably: weakening the claim. Confidence of a problem	The user is unfamiliar with the application domain	A UP is a problem. A UP is <u>not</u> a problem

Table 9 Reasoning patterns when rating UP confidence level

Assessing own confidence	**Frequency**
Experiences of the problem	5
Understanding of the problem	6
Proposing a problem solution	3
	14
Assessing user or other values	
Estimating user experiences	3
Estimating priority to fix the problem	3
Estimating problem severity	4
	10

Table 10 Reasoning patterns for whether two UPs were similar or not

Reasoning patterns on consolidation	**Individual filtering**	**Collaborative merging**	**Total**
Looking at tasks, context or otherwise referring to user	6	3	9
Characteristics of problem - Criteria for extraction or severity	1	2	3
Reading text descriptions	5	0	5
No articulation of decision	4	3	7
Total:	16	8	24

Table 11 Severity - Argument model template

Data	Warrant	Backing	Qualifier	Rebuttal	Claim
Facts from the evaluation protocol about task completion, time, user behaviour, need for fixing, characteristic and background of users	A rule describing a condition for UP severity with reference to task completion, time, user behaviour, need for fixing, characteristic and background of users	Definition of Severity, Software development practices, Human Computer Interaction	Uncertainty of severity rating or evaluator had the same experience	The severity may depend on individual users	UP is of a certain severity, minor, moderate or severe

Table 12 Reasoning patterns on severity rating

Reasoning patterns	Individual extraction	Collaborative consolidation	Total
Looking at tasks, context or otherwise referring to user	19	3	22
Characteristics of problem - Criteria for extraction or severity	1	0	1
Reading text descriptions	2	0	2
Discussions on severity rating	0	3	3
No articulation of decision - No answer	2	2	4
Total:	**24**	**8**	**32**

Biographies

Ebba Thora Hvannberg is a professor of Computer Science, School of Engineering and Natural Sciences, University of Iceland. Her research areas include human computer interaction and software engineering. The main focus has been on methods for usability evaluation in various areas such as e-learning management, air-traffic control and crisis management.

Effie L-C Law is a full professor in Human-Computer Interaction, University of Leicester, UK. Her main research interest is usability and user experience methodologies, which are applicable to various domains, including technology-enhanced learning, healthcare, and cultural heritage. Her current focus is on automatic emotion analysis with multisensory data and on the relation between emerging technology and wellbeing.

Gyda Halldorsdottir is a freelance working project manager, University of Iceland and University Hospital of Iceland. Her research areas include human computer interaction and health informatics. The main focus has been on usability evaluation, air traffic control, crisis management and data management.