The Influence of Domain Expertise in Usability and Information Architecture Evaluations

Mathias Pedersen & Anders Østergaard Bertelsen

10th Semester Master's Thesis, 2015

MSc. Information Architecture, Aalborg University

Supervisor: Mette Skov, Aalborg University

AALBORG UNIVERSITY
DENMARK

The Influence of Domain Expertise in Usability and Information Architecture Evaluations

10th Semester Master's Thesis

MSc. Information Architecture

Mathias Pedersen

Anders Østergaard Bertelsen

Supervisor: Mette Skov, Aalborg University.

Submission date: June 1st, 2015.

Character count: 313.906 (130,8 standard pages).

Abstract

The purpose of this thesis is to test the hypothesis stating; *that since users with domain expertise browse, perceive and use domain specific systems differently than domain novices, involving domain novices into the evaluation process would help to locate and improve more disadvantages that would be overlooked, but also yield more results about the advantages of the system that are less focused on content, but the overall usability aspect instead*. The hypothesis has been compiled from the literature on the subject of usability evaluations. We have chosen this hypothesis, as we want to contest the prevalent importance of using participants with domain expertise (intended users) in usability evaluations, as we in our studies in Information Architecture have experienced that novice test participants can contribute to usability evaluations too.

The first part of the thesis consists of a presentation of the general theory on the subject of usability evaluation and Information Architecture, which also covers the literature stating the importance of using intended users as participants. This is followed by recent research studies on the subject of domain experts compared to novices in usability testing, as there is uncertainty on when to use experts or novices as usability test participants.

The test methods used to test our hypothesis are Think-Aloud and Card Sorting, which are being utilised on twenty participants, who consist of ten domain experts and ten domain novices. The experts were chosen from their expertise on mountain biking and the novices were chosen at random, so to study the differences between a carefully selected group of participants and a group consisting of participants chosen at random. The mountain biking domain is chosen, as it fits the context domain of the websites used for the usability tests.

The thesis is concluded with our recommendations for the choosing of participants in usability testing. The recommendations are based on our findings in the analysis of the data provided by the two participant groups in the usability tests, and reflect the results of when we found the use of domain novices to be as useful, or more useful, than using domain expert participants.

The conclusion shows that the participants with no domain knowledge can in some aspects of the usability test be useful, and it will include our recommendations for what future usability and Information Architecture evaluators could consider, when completing their research design.

We would like to express our thanks to Mette Marie Kronborg and the rest of Feriecenter Slettestrand, our supervisor Mette Skov, and all of the test participants for all of your help in the process of completing this thesis.

1.0 **Introduction** 8

2.0 **Theoretic Background** 13

3.0 **Methods** 29

- ❖ Heuristic Evaluation 31
- ❖ Think-Aloud 32
- ❖ Card Sorting 37

4.0 **Research Design and Data Presentation** 47

5.0 **Analysis** 87

- ❖ Usefulness, Effectiveness and Efficiency 90
- ❖ Learnability and Satisfaction 101
- ❖ Errors and Safety 113
- ❖ Labels 116
- ❖ Navigation 129
- ❖ Discussion and Summary 137

6.0 **Conclusion** 142

7.0 **References** 146

Structure

Introduction: The introduction includes our motivation, the context for the thesis and our research questions and a case description.

Theoretic Background: Includes a definition of *usability*, aspects on usability testing in general and the theoretic understanding behind participant domain expertise.

Methods: The methods section includes descriptions of the theory behind the methods that we will use in this thesis (Heuristic Evaluation, Think-Aloud, and Card Sorting) together with a description of what domain expertise means in this context.

Research Design and Data Presentation: A description of the practical aspects of our research design, and a presentation of the data collected during the usability tests.

Analysis: An analysis of each usability attribute, and how the data for each attribute is being affected by test participant domain expertise. It also includes a discussion on some of our choices during the analysis.

Conclusion: An overall conclusion of this thesis and our findings, combined with our recommendations for future evaluators.

Appendix Overview

This is an overview of the appendixes that will be referred to during the entirety of this thesis (file types and descriptions).

The appendixes can be found on the enclosed SD memory card.

01. [PDF] Card Sorting Tasks for Test Participants.
02. [DOCX] Cards used for Card Sorting (A and B cards).
03. [PDF] Think-Aloud Tasks and Context for Test Participants.
04. [XLS] Four Card Sorting analysis spreadsheets containing:
    a. Domain Expert A-cards.
    b. Domain Expert B-cards.
    c. Domain Novice A-cards.
    d. Domain Novice B-cards.
05. [DOCX] List of test participants (number, age, sex, domain expertise and median ages).
06. [DOCX] Notes for Think-Aloud video files (includes time taken for tasks and number of mouse clicks).
07. [PDF] Card Sorting data showing labels chosen for front page, unnecessary labels and labels that were used for headlines for categories.
08. [PDF] The results of the Heuristic Analysis session.
09. [MOV] Videos from the Think-Aloud tests.

1.0 Introduction

Information Architecture deals with different understandings of the same system, often visualised by dividing the system into three separate, but co-dependent parts: the context, the content and the users of the system, often called the three circles of Information Architecture (Morville & Rosenfeld, 2006, p. 25). This model is a simplified model that helps to show how systems, such as websites, are not static constructs, but are organic in the sense that they are constantly evolving to attend the needs of each part of the model. The understanding of Information Architecture, especially good Information Architecture, has to incorporate each of these aspects in order for them to influence the design in a way that is informed by all three aspects.

One of the tools to validate if this is the case with a system, such as a website, is to complete an evaluation of that system. The process of designing a website should take the target group of the intended website into account, to make sure that the design complies with the behaviour and preferences of the users in that target group. This needs do to ensure that the users will take the website to use, without feeling bothered by the system or confused by the way the website content has been organised by the designers, as there might not be compliance with how website designers and the users of the website might view and understand the website and its contents. The aim is to create a system or website that is enjoyable and satisfying for the users to use, in situations and contexts where it is appropriate. The problem often arises from the fact that frequently, very few people design most websites and services, but that these design products have wider user populations that may have been somewhat overlooked, which can result in designs that are designed only for the designers, and not the users (Rogers et al., 2011, pp. 433-434).

As Information Architecture considers these designs organic and in constant change, it is therefore appropriate to reflect on them during the design process. This does not only apply to when a website concept is being developed and is still in its first iterations of the design process, but also applies later in the design process when an already established website is being upgraded or changed in a way that interferes with the use context, the content or the users. By doing so, new information about changed use requirements can be obtained in the evaluation process, and this newly gained information can be embodied into the upgraded or changed design to the advantage of both users and designers.

Therefore, it is very important to consider how to complete the evaluation process, and what type of results are desired. These variables include, but are not limited to, what types of tests that need to be completed, what needs to be tested for and when and where to incorporate users in the evaluation

process. Heuristic Evaluations are completed solely by experts, for example by reviewing a website's navigation and interface, and comparing the results to accepted usability principles, where other types of evaluations include users in the evaluation process. For example, this can be done by completing tests that yield results about how the users prefer the website's usability and navigation to be designed, or by investigating how easily the users manage to complete tasks by using a website, and reflecting on where and how problems can be solved to make it easier or less bothersome for the users.

However, as suggested earlier, there are big differences in how users think, understand and behave on websites. Rogers, Sharp and Preece (2011) describe how users have individual and developing mental models of a system, which forms their behaviour and use of a system different on an individual level (p. 86). Users with good mental models of a system will be able to recover more easily if something goes wrong with their interaction with the system, where other users with poor mental models might give up (Benyon, 2010, pp. 32-33). Users' mental models can be improved by interacting with the system, or systems like it.

How wide the range of user populations within the target group of a website or system is, depends very much on the target group itself. Is the system aimed to please a very wide target group of people, or is it targeted for a group of users that have a specific domain knowledge in common? This is an important question to take into account, when choosing what type of product is being evaluated, and which users should be used in the user tests of that evaluation. Much literature on user testing principles suggests that when conducting user tests for evaluations, it is necessary to condensing the test users to the core target group of the website or system that is being evaluated (Bednarik & Tukiainen, 2005; Botella et al., 2014; Dou et al., 2009; Karapanos et al., 2008; Kinney et al., 2008; Kjeldskov et al., 2010; Lazonder et al., 2000; Nielsen & Molich, 1990a; Nielsen, 1992; Nielsen, 2000). The argument is that condensing the test users to match the target group helps to understand and evaluate the website or system from the best possible perspective, as it aids to eliminate redundant information and not provoke or incite usability or user experience problems that the target audience would be less likely to experience. However, Information Architecture theory is built on understanding the same problems with different perspectives, seeking to eliminate existing problems, but also future problems that might not be relevant now, but could become relevant in the future, as the system develops. "*No single approach can stand alone as the one right way to learn about users and their needs, priorities, mental models, and information-seeking behavior. This is a multidimensional puzzle—you've got to look at it from many different perspectives to get a good sense of the whole*." (Morville & Rosenfeld, 2006, p. 247).

Our assumption is, that conducting an evaluation by looking at the perspectives of different user types within the same target group can help to improve websites or services where there is not only *one* target group, but several target groups within the same context. We aim to investigate how evaluation results from user test groups of users that have domain expertise matching a website's content domain differentiate from the evaluation results from user tests with users that are domain novices. The idea is that even though some test users might not match the domain of a website perfectly, but rather fit into a more generic user group, their test results might yield useful information on how to prevent implications that test users with a stronger domain knowledge would overlook or ignore because of different browsing habits.

1.1 Research Context and Questions

Conducting evaluations of systems or websites is an excellent way of understanding how that system is perceived by its intended users. Evaluations, whether they are based on expert heuristics or substantiated by involving the users, helps to understand how that system could be designed, changed or developed to suit the needs of its end-users.

When surveying the theory concerning evaluations that involves users, it is often suggested to make sure that the test population consists of users with a strong domain knowledge, in order to make sure that the evaluation results are within the boundaries of that exact target group. But not every system, website or information architecture has only *one* target group, but several target groups within the same context. In such cases, we aim to investigate how evaluation results from user test groups of users that have the domain knowledge matching a website's content domain differentiate from the evaluation results from user tests with users that are domain novices.

Our hypothesis is, that since users with domain expertise browse, perceive and use domain specific systems differently than domain novices (often with more ease and by utilising more content), involving domain novices into the evaluation process would help to locate and improve more disadvantages that would be overlooked, but also yield more results about the advantages of the system that are less focused on content, but the overall usability aspect instead.

- How does usability and Information Architecture evaluation results from users with domain expertise differentiate from evaluation results from users with no or little domain expertise?
- How should these results be considered in future evaluation research designs?

1.2 Case Description

To elucidate this hypothesis, we have entered into collaboration with Feriecenter Slettestrand. They accommodate holiday visitors who can be divided into four overall groups of visitors: families, seminar participants (municipalities, firms etc.), people with disabilities (for example wheelchair users) and mountain bikers. The reason for Feriecenter Slettestrand being popular for mountain bikers, among other things, is through the location of the holiday centre, as the nature in which the centre is located, is a combination of hills and forests, which have made Feriecenter Slettestrand one of the most popular visiting locations for mountain bikers from around Scandinavia. Their 22 kilometres mountain biking track, which has been awarded as the best in Denmark[1], is part of the surrounding terrain, and they are already planning to expand it to a 30-kilometre track. The holiday centre is a family owned business and the family itself are very much a part of the Danish mountain biking community[2], which furthermore have made them a part of a network of people with the interest of mountain biking.

The surrounding nature and focus on pleasing the mountain biking community have also resulted in mountain bikers being a large part of the visitors who visit Feriecenter Slettestrand. The interest in mountain biking is also shown through their current website, where the mountain biking part of their website is frequently visited, as shown by Google Analytics reports. The website was developed during 2011, but the website has rarely been updated on a regular basis, which has left the website in a state, where much of the content has been neglected and the amount of content reached a point where Feriecenter Slettestrand wanted to start it afresh. At the beginning of this development process, they reached an agreement on wanting to further gratify and interact with their mountain biking visitors, as to further encourage their mountain bike users and network to visit them and make use of their mountain biking track. This has resulted in Feriecenter Slettestrand now being in the process of developing a standalone website for mountain biking, in which they plan to update the content more frequently. Through this website, they will keep news and activities up to date and they hope that this will generate a bigger interest in their holiday centre. The mountain biking website will link to their main website multiple times throughout the website and through that, they also hope to persuade more people to visit their main website and book holiday stays.

[1] http://www.vorespuls.dk/artikel/top-3-danmarks-bedste-mtb-ruter
[2] http://www.nordjyske.dk/nyheder/han-er-aarets-mtb-rytter/ddd0834a-53bb-408c-869d-d3c9c0cab272/112/1513

We have entered the process in the middle of the development of the two websites (the main website and the mountain biking website), where the overall wireframes and ideas for content have already been developed. The development is a collaboration between Feriecenter Slettestrand and a website designer[3], who has been their partner since the first website was designed. The designs and wireframes for the old and new websites have not had much user testing or other kinds of evaluations done, and they are therefore primarily a result of the wishes from Feriecenter Slettestrand and the expertise of the website designer.

To sum up, Feriecenter Slettestrand currently have a website, which they are in the process of replacing with an updated main website for presenting their holiday centre, restaurant, accommodations, activities and booking information, and a new standalone website addressed for mountain biking news and activities, serving a user group with a more defined domain expertise within the field of mountain biking.

This iteration is the step before the initial implementation of the new websites and the replacing of the old website. We will therefore be in a position, where we can test and evaluate the new mountain biking website, and the existing main website at a point, where they are almost ready for implementation of the mountain biking website. We do not have an opportunity to test or evaluate the new main website, but we are able to evaluate the existing site, and use it as a context for this thesis.

This case is interesting for us, as it deals with a very specific user group for the evaluation, which our preliminary interest is conceived from: The testing and evaluation of a website, where the users can be pinpointed as people being interested in vacation, where either mountain biking, or the limitation of being disabled, put these people into specific user groups with different kinds of domain expertise. Our focus will largely be on the mountain biker group, as the standalone website developed for Feriecenter Slettestrand is targeted at mountain bikers, making the user segment for that website even stronger. This makes the Feriecenter Slettestrand case interesting for our hypothesis, as the user group is a specific user group having knowledge on the domain of mountain biking, which greatly separates the mountain bikers from the general users, as it is necessary to understand the language of mountain biking, as to understand certain aspects of it (technicalities of bikes, courses, phrases etc.).

[3] http://www.mindthemedia.dk/

2.0 Theoretic Background: Introduction

Evaluating websites is a process involving many factors, depending on the type of evaluation and the methods used for completing the evaluation. In this section we will expound on the context of this thesis, describing the type of evaluation we chose for testing our hypothesis, and also explain the theoretical foundation of which our hypothesis was composed. We will also enlarge on the term *usability*, as the evaluation we work with is primarily based on assessing the usability aspects of Feriecenter Slettestrand's websites, and how that usability evaluation can work as a research context for our hypothesis.

Later in this section, we will focus on the *user* aspect of the websites being evaluated, and how these users are important for the design and evaluation process in usability studies. We will also enlarge on the importance of being able to, from the designers' perspective, differentiate and classify users in classifications with dissimilar characteristics, and how we do this in our evaluation process, and why.

2.1 Defining Usability

As earlier described, this thesis will evolve around a usability evaluation of Feriecenter Slettestrand's websites, and be used as a context for testing our hypothesis. Therefore, it is necessary to include a definition of what usability is, and understand the different ways that usability can be evaluated on for example websites.

In short, usability focuses on how well users can learn and use a product or service, and how this product or service can help the users to achieve their goals, whatever the goal might be. It also focuses on how satisfied the users are with this process (U.S. Department of Health and Human Services, 2015b). This sounds simple enough; create a product or service that is tailored to aid a specific user goal, and you are set. Basically, that is the overall idea of usability, but in reality usability is a much more nuanced and intangible process that requires an understanding of the use context and the people ending up using the product or service, and how they think and react to certain elements of the design. Therefore, usability applies to every aspect of a system, product or website where there might be user interaction (Nielsen, 1993, p. 25), and finding aspects of a website that does not include human interaction is, of course, very difficult as most websites are made for interaction in some way or another.

As more and more complicated and complex computer systems have found their way into our everyday life, the need for being able to understand and use these systems has grown. Because of this

tendency, usability has become extra popular since it started being more relevant when computers, and use of them, became more widespread. Since then, usability has become more and more relevant, as the use of other electronic systems and products has grown and keeps growing (think VCRs, pagers, smartphones, laptops, Smart TVs, smartwatches etc.).

Subsequently, the interaction process between the product and the person(s) using it is where usability is important. The interesting question is now, how then can usability be measured, and how is something *usable*? The International Organization of Standardization defines usability as "*The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.*" (ISO/IEC, 1998). This definition includes factors such as *effectiveness*, *efficiency* and (user) *satisfaction.* There are other variants of this definition though; usability expert Jakob Nielsen describes usability as associated with five factors (or attributes): *learnability, efficiency, memorability*, *errors* and *satisfaction* (Nielsen, 1993, p. 26), whereas Jeff Rubin and Dana Chisnell define usability by attributes of *usefulness*, *efficiency*, *effectiveness, learnability, satisfaction* and *accessibility* (Rubin & Chisnell, 2008, pp. 4-6). Rogers et al. also include attributes such as *safety* (the product being safe to use) and *utility* (the product having good utility) (Rogers et al., 2011, pp. 18-20). Sometimes, some of these attributes are summed up as the simpler term *intuitive design* (U.S. Department of Health and Human Services, 2015b).

Information Architecture is also very relevant in usability evaluations, and for the usability definition in general. Information Architecture adds to the definition, by taking into consideration other relevant attributes such as navigation systems, organisation of content and how this content is labelled – all attributes that can affect or change how the usability of a website or system can be measured or how it is being perceived by users. In the forewords of Morville and Rosenfeld's "*Information Architecture for the World Wide Web*", Jakob Nielsen formulates the connection between the two disciplines as such: "*Usability is an important, though not the only, determinant for the success of a web site or an intranet. Information architecture is an important, though not the only, determinant for the usability of a design. There are other issues, but you ignore information architecture at your peril*." (Morville & Rosenfeld, 2006, p. xi). Information Architecture is important to shape and create information systems or websites that supports good usability (Morville & Rosenfeld, 2006, p. 4). In a way, the disciplines are almost mutually exclusive, meaning that without some sort of Information Architecture foundation, it is hard to improve many of the attributes of the usability term, as they are closely connected.

Because of this variety of different attributes that have been added to the usability term, the definition of usability is a bit subjective, depending on how the different usability attributes are being valued important on the system or product in question. In our evaluation, we will primarily focus on:

Effectiveness, usefulness, and efficiency: The system allows users to achieve their goals, if the goals are supported by the system. For example, a website for comparing flight ticket prices might be useful if the goal is to only compare flight prices, but in such cases, most users would probably also prefer to be able to buy the tickets. The system should be efficient so that when the user has learned how to use it, it can be utilised to achieve a high level of productivity (Nielsen, 1993, pp. 30-31). Usefulness also influences the user's willingness to use the system at all (Rubin & Chisnell, 2008, p. 4; Rogers et al., 2011, pp. 19-23). Why would a user compare flight ticket prices on a site that does not allow buying tickets, if the end goal that he wants to achieve is to buy a ticket at the lowest price? It would be most sufficient if the user would be able to compare the prices, and buy the tickets directly at the same website, allowing him to achieve his goal quickly.

Learnability and satisfaction: The ease of learning on a system is important, especially to new or first-time users that might never have seen the user interface before. It is necessary for the system to be tailored to fit the user's level of knowledge of the system's interactive functions, so that the user can quickly learn and use the system. That does not mean that every system should be easily managed by every possible user, but it should be fit to the specific target group of the system. Some systems might require intensive training before use, but if that fits the target group, and the users expect this to be the reality, that is completely fine (Nielsen, 1993, pp. 27-30; Rubin & Chisnell, 2008, p. 4).

If the learnability curve is adaptable by the users within the target group, it can also help to heighten the overall user satisfaction on that system. The satisfaction refers to the users' perceptions, opinions and feelings toward the system, and a higher user satisfaction leads to better user performance, which means that the users are more prone to achieve their goals, than if they were using a system that did not provide user satisfaction (Rubin & Chisnell, 2008, pp. 4-5).

Satisfaction can be difficult to measure in usability evaluations, since it is hard to derive from a test participant's actions if he is satisfied with the system that he is using. For measuring satisfaction somewhat consistently, as it is a very subjective attribute, Nielsen recommends using short questionnaires or to conduct retrospective interviews with the users, after they have used the system for a real task (Nielsen, 1993, p. 34). Another way is to focus on a concurrent style of Think-Aloud tests, where you allow the participants to verbalise thoughts and reasons for their behaviour out loud,

while completing a set of planned tasks. The advantages and disadvantages of concurrent and retrospective Think-Aloud methods are further discussed in the Think-Aloud method section 3.2.

Errors and Safety: Of course, the system should have a low rate of errors. This helps to ensure that users also make fewer errors during use of the system. A system error is not necessarily equivalent to something on the system not working or breaking during unexpected use, but can also be instances of users doing something wrong, when trying to achieve a specific goal (Nielsen, 1993, pp. 32-33). System errors, or catastrophic errors, is not something that users necessarily can prevent – but they can find and experience them. These types of errors should obviously be removed or minimized in frequency. The safety attribute involves making sure that users do not end up in undesirable situations, which they might have a hard time recovering from, or make errors accidentally (Rogers et al., 2011, pp. 19-21). An example of this could be in a software interface where the "save" button is placed directly next to the "delete" button, making it easier to miss the right button, and click the wrong one which deletes the data that the user wanted to save, rather than clicking a less critical button that could have been placed between the two.

Intuitive Design: Intuitive design is "*a nearly effortless understanding of the architecture and navigation of the site*" (U.S. Department of Health and Human Services, 2015b). This usability attribute is much related to Information Architecture attributes dealing with navigational aspects of the system's interactive user interface, which has a large influence on how, and how easily, users explore, find and navigate information that is needed to achieve their goals (Kalbach, 2007, pp. 34-37). Letting users know how to get to where they want to go also helps boost effectiveness and efficiency (Reiss, 2012, p. 4).

Usability as a term consists of many different attributes, that all define one aspect of usability. There are many aspects, depending on the type of product usability is being ascribed to. From one perspective, usability can be summed up as making sure that there is an "*absence of frustration*" (Rubin & Chisnell, 2008, p. 4) in using that product. These attributes are also very dependent on the context of use, what type of users are in play and what their goal is (Barnum, 2011, p. 11).

In addition, it is worth noticing that usability differs from things that are "user friendly" (even though "user friendly" was the term of choice when personal computers started to gain popularity). As Nielsen puts it: "*users don't need machines that are friendly to them, they just need machines that will not stand in their way when they try to get their work done*." (Nielsen, 1993, p. 23). He also

mentions the importance of considering users' different needs, and what might seem "friendly" to one user might feel tedious to the next. We will elaborate on how to differentiate users in section 2.6.

2.2 Evaluating Usability

Now that the definition of usability has been cleared up, it is necessary to examine how usability can be measured. To evaluate the usability on a website, system or product, *usability testing* as a research tool can be used to indicate which of the usability attributes should be improved on, and which are working well for the users. Usability testing, as a tool, can consist of many different activities and methods that help the designer improve the existing product, by knowing where to improve it in order to raise the level of usability. In usability testing, the primary goal is to determine how typical users (typical users in the defined target group) solve representative tasks (Rogers et al., 2011, p. 438). For example, when evaluating a website it would be relevant to use usability testing to evaluate user interfaces, navigation, labelling and other design elements, and see if these elements are working in the way they are intended with that target group, from the designer's perspective. Some of the ways of measuring this includes recording the number of errors users make, and comparing the number to how it could have been done, and how much time the users take to perform their tasks. As mentioned in section 2.1, other usability attributes, such as user satisfaction, can be measured by interviews or questionnaires that inquire into these topics.

2.3 Usability Testing Methods

Most of the methods for evaluating usability require a group of representative participants, who are used for measuring performance and efficiency (Rogers et al., 2011, pp. 433-434), but there are also methods that do not; for example Heuristic Evaluations. Heuristic Evaluations are conducted by usability experts in order to detect usability issues on a system design, often before participants are introduced in other types of evaluation methods.

Some of the often used methods for evaluating usability in usability testing processes, which also include participants are, as already mentioned, interviews or questionnaires. Arguably one of the most valuable and used methods for evaluating usability is the Think-Aloud method, which involves participants that "think aloud" when they are using the system that you want to evaluate, enabling the designers to get insight into what thought processes the users go through when using the product being tested. Other popular methods include focus groups, observation, logging, eye-tracking, user feedback, walk-throughs, paper prototyping and Card Sorting (Bødker et al., 2008, pp. 243-335; Nielsen, 1993, pp. 207-225; Rogers et al., 2011, pp. 437-443; Rubin & Chisnell, 2008, pp. 16-20).

We primarily use Think-Aloud and Card Sorting for testing our hypothesis, and will elaborate on these methods later in the methods section.

2.4 Types of Usability Evaluation Settings

Another aspect of usability evaluations that is important to consider is the setting in which the testing takes place. The setting can help to adjust the level of environmental control, and is generally dependent on what is being tested (Rogers et al., 2011, p. 436). Rogers et al. (2011) have described three broad categories of evaluation settings, which they have based on the level of control and user involvement (pp. 437-443):

1. **Controlled setting**. The controlled setting includes participants, and includes tests that require some control of what is able (or rather, not able to) influence the participants and their behaviour. The setting can for example be a usability laboratory designed to create and control the ideal testing environment – it can be quiet, provide space for observers without disturbing the participants, and include special equipment for special types of tests, for example eye-tracking equipment (Barnum, 2011, pp. 26-30).
2. **Natural setting**. This setting also involves participants, but not provide many, if any, options for controlling the test environment. Instead, this setting is excellent for testing how products would be used in real life scenarios or "in the wild" (Rogers et al., 2011, pp. 440-441).
   In comparison to the controlled laboratory setting, it is harder to anticipate and be present when something valuable to the evaluation happens, but it can provide better results on how the actual use context of the product influences user behaviour (Barnum, 2011, pp. 38-41; Rubin & Chisnell, 2008, pp. 100-101).
3. **Any setting not involving participants**. The last test setting does not involve participants, but is conducted by using researchers and experts within the field that is being evaluated. An example could be Heuristic Evaluation where experts evaluate and identify usability problems, or walk-throughs where consultants recreate the steps users would have to go through, based on predicted user behaviour (Rogers et al., 2011, pp. 441-442).

For evaluating usability of Feriecenter Slettestrand's websites, we are mainly using a controlled setting (not a formal usability laboratory, but rather informal laboratories (Barnum, 2011, pp. 37-38) where we bring our own equipment, but still have the possibility of taking notes, recording the participants' actions and thoughts (when they are thinking aloud) and letting them work in a quiet

environment), and a setting not involving participants (for the Heuristic Evaluation).[4] These settings are sufficient for our evaluation process, as the websites we are testing are not meant to be used in a specific setting, location or "in the wild", and conducting the evaluations with participants in a controlled setting allows us to make sure that the participants focus on the websites and their content, rather than having to worry about the environment around them and if that environment would influence the data from the evaluations (Nielsen, 1993, pp. 205-206).

When the setting and method aspects of the usability test have been chosen, and there are sufficient participants to help with the evaluation, the usability evaluation can take place. The reason for completing such an evaluation in the first place is to collect data on performance and preference measures from the participants (Rubin & Chisnell, 2008, p. 25). This data is important for producing the next step in the evaluation of the product; a recommendation on how to improve it. This can include proposals for redesigning specific areas of the information architecture, navigation, labelling etc. It can, and should, also include both the advantages and disadvantages that have been found during the usability evaluations, so that the evaluation is constructive criticism, rather than only a critique of the disadvantages or usability problems.

2.5 Limitations of Usability Testing

Even though the settings of our usability evaluation are not the most influential factors of our study, it is worth noticing that one of the limitations of usability testing is that it is very hard to simulate the appropriate context for the product that is being tested. A controlled laboratory (formal or informal) will never simulate the same conditions that the product will be used in, in real world scenarios. Testing will always be an artificial situation, even when testing in the wild or natural settings, and will always only be a depiction of usage (Rubin & Chisnell, 2008, p. 26).

We have discussed which usability attributes could be measured in usability testing (effectiveness, efficiency, satisfaction etc.), but in a review of 180 studies by Kasper Hornbæk (2006) where the common practices in measuring usability are discussed, categorised, and critically reviewed, it is concluded that there are evident problems in how these types of attributes are being measured (Hornbæk, 2006). One of the suggestions is to make a clear distinction between *subjective* and *objective* usability measures (based on if the usability measure is based on users' perception or

[4] We do, however, have experience conducting Think-Aloud tests in natural settings from an earlier semester, where we evaluated the usability of a mobile application (Housing Enabler) that was designed to be used in work situations, which required the user (occupational therapists) to walk around houses (both indoors and outdoors) while using the application on a smartphone or tablet.

attitude toward it, or if it is not dependent on the user at all). The argument is that such a distinction would aid in the selection of usability measures to assess, and that having both subjective and objective measures could help usability studies review usability from different perspectives.

Even though different perspectives are implemented in the evaluation process, test results do not prove that a product works in general (Rubin & Chisnell, 2008, p. 26). Even though some aspects of the product might have excellent usability, and the data collected from the tests show that the participants find the usability to be satisfactory, it does not take into account the whole of the product.

Many usability tests are also completed with low sample sizes that yield qualitative results, and with participants that in many cases are not completely representative of the target group of the product being evaluated. The actual end users of a product can be hard to identify and define, and recruiting participants that match the actual end users can in many cases be problematic, especially when using a small sample group of participants (Rubin & Chisnell, 2008, p. 26). We discuss this further in the next section.

2.6 Differentiating One User from Another

An important part of completing an evaluation process, where the system being evaluated and developed on is designed for someone else than the designers themselves, is to implement a controlled setting where the evaluators and designers can control or eliminate outside influences and distractions, and use this setting to involve users in their design process (Rogers et al., 2011, p. 438). As explained in the previous section, usability testing is such a process, where the setting can be organised to be a controlled or natural setting, depending on what is deemed necessary for that stage of the evaluation. As we see it, one of the most important goals for evaluations is, however, to experiment with the selection of users used for evaluation. For usability testing involving any users, it would be obvious to investigate how typical users of the system being evaluated, react and behave on the system, and see if they can complete the tasks that the system was designed for. The evaluation process can help to investigate where in the system's design the users encounter factors that hinder or slow their task completion process, allowing the evaluators to observe *why* this happens.

Why something happens to the users in the evaluation process, is what helps designers understand the metrics of that advantage or disadvantage happening, and learning how it can be utilised or changed in the system to increase the usability of the system, and the use experience by the future users. It proves a little more difficult than it sounds though, as understanding *why*, the designers and evaluators are required to understand how problems appear, and why they are a problem in the first place. In

usability studies, this could often be the information architecture giving rise to problems within for example navigational aspects of the system. Observing users perform tasks and make decisions in an evaluation setting shows more about user interaction with the system, than what most designers would have been able to derive from reports or presentations from non-users (Rogers et al., 2011, p. 438). Another reason to complete evaluations in this matter, is to include several users for the sake of variety. As Morville and Rosenfeld state, no two users think, act, read or behave the same (Morville & Rosenfeld, 2006, p. 4), and so usability testing involving several users can help draw a larger picture of how users understand and use a given system from a qualitative standpoint. Their individual mental models of how a given system or website should operate or work (what should happen when a button is clicked, the general layout of menus and content, where to expect finding particular information etc.) are constantly developing and changing by interacting with like websites or systems, where they see how their actions are tied to certain behaviour of the website (Benyon, 2010, p. 32; Wills & Hurley, 2012).

Because of this reason, it is important for designers to be able to understand how users can be differentiated, but also where they can be classified in groups, which is important for the developers as they need to be able to classify different user groups within the overall target group of the website or system (Morville & Rosenfeld, 2006, p. 235).

One of the reasons for classifying users in user groups is that it is unreasonable to design a system or website that takes into account every conceivable usability problem (or other type of design problem). It is, however, much easier to manage few groups of users that have been grouped together because of properties that tie them together in certain ways that it helps to differentiate between the groups, when designing for them. There are many variables, which can be used for this. In quantitative surveys, variables such as demographics (for example sex, age, income, and level of education) are often used. In this thesis, the approach is qualitative as we will focus on data from individual users and their perspectives, and therefore classify the difference in users on a more general aspect; domain expertise.

In *Usability Engineering* (1993), Jakob Nielsen argues that the most important issues, when dealing with usability, are the characteristics of the users, and the tasks they perform (Nielsen, 1993, pp. 43-44). He also argues that it is likewise important to emphasise, that even though the target group that is being designed for has several different groups of user classifications, based on their characteristics, it is still possible to design a system or website that is good for many or all of these classifications, if

there is a way to include usability features that take into account the aspects from each classification's evaluation results.

Nielsen continues defining users with different levels of expertise, often referred to as either novices or experts (Nielsen, 1993, pp. 43-48). This classification is one of the most commonly used, and it is a resemblance of the users' experience with a given system, website, user interface or domain that defines them as either novices, experts or something in between the two. Nielsen's definition of novice and expert users (Nielsen, 1993, pp. 43-48) is based on the user's experience with computers (or other relevant technology, as the 1993 definition can be applied to a lot more than only (desktop) *computers* now (think of smartphones, tablets, watches etc.)), and their knowledge of the domain in question. For this thesis, when mentioning users, we will often be referring to the last part of the definition, the domain expertise of the users within a specific domain. Russell-Rose and Tate's definition of domain novices and experts is based on the same metrics as Nielsen's, but they updated the definition from *experience with computers* to what they simply call *technical expertise* (Russell-Rose & Tate, 2013, p. 4).

What then defines *domain* expertise? Russell-Rose and Tate have an example where they ask people how they are most comfortable taking photographs – do they prefer point-and-shoot cameras/their smartphone for quick and easy photographs, or are they comfortable with using a SLR[5] camera that has many buttons, dials and lens options for fine-tuning picture quality, focus, exposure and aperture? There is no wrong or right, but whichever is preferred, is an indication of how comfortable each option is to the user, also indicating their expertise within the *domain* of photography, where novices might gravitate towards the quick and easy solution, and people with better domain knowledge might gravitate towards the SLR camera (Russell-Rose & Tate, 2013, p. 3).

If the SLR camera had been the context, it would have rather been a context based on *technical* expertise, and people would have been asked how comfortable they are with one type of camera – the SLR. In such a case, some users with none or little technical expertise would most likely use the camera's "auto-settings" mode (even though they might have preferred a point-and-shoot camera) whereas technical experts would probably use the full manual mode because the know the functions of the camera, and how they should be utilized in which settings.

[5] Single-Lens Reflex camera or Digital Single-Lens Reflex camera (DSLR).

However, domain knowledge does not have to be about photography. Whatever the domain context is, if the website or system being designed has users with varying level of expertise, it is necessary to differentiate between them in usability tests, and use the test results to strike the right balance between domain contexts that are too vague or too complicated.

To sum up, there are two types of expertise that can have an effect on how users seek and find information: domain and technical expertise (Russell-Rose & Tate, 2013, p. 4). Technical expertise is an indication of how well the user understands and is able to use the technologies that are relevant for the evaluation. This includes their proficiency in using the internet, computers, search engines, SLR cameras etc. Technical experts would be defined by having excellent technical skills, whereas technical novices would have a hard time using the relevant technologies on their own hand, and might need instructions or training first. Domain expertise holds the same premises, but instead of being assessed on their skill using the technologies in play, they are assessed on their familiarity with a given subject domain (photography, mountain biking, the medical industry etc.).

Both types of expertise are user characteristics that can be used for classification of test participants in usability testing. There can be four types of users, dependent on their expertise within domain and technical dimensions (Russell-Rose & Tate, 2013, p. 4):

- Double experts
- Domain expert/technical novices
- Domain novice/technical experts
- Double novices

As we will later explain further, we are not very concerned about the technical expertise of the participants we are using in our evaluation process, as we are evaluating websites. Websites, and the internet in general, has been around for so long that 93% of all Danish families (statistics from 2014) have access to the internet from their homes (Danmarks Statistik, 2014), which is why we consider the technical expertise in this case to be almost irrelevant. We are also only using participants that have used the internet before for our usability tests.

2.7 Experts and Novices

Differentiating between novice and expert users is not something new. In his 2000 article, Jakob Nielsen describes how *web usability*, with the focus on increasing ease of learning for novice users has shifted its target multiple times throughout the last handful of decades. He describes it as a

pendulum that swings back and forth between trends. At some points in time, the focus on usability was primarily with domain expert users as they were basically the only group using this kind of technologies that included usability, but other times, for example when Apple introduced the personal computer and more people began using different types of graphical user interfaces, the pendulum swung back to focus on the novice user (Nielsen, 2000). In the article, he also predicted (correctly)[6] that "*There will likely be huge growth in Internet-based applications that are not really websites but where users perform daily tasks across the Internet. For example, online calendars and maybe even entire office suites*." (Nielsen, 2000). However, the type of novice and expert users Nielsen use examples of in this article is novices and experts based on their ability to use a specific website or application, by having spent various amount of time with it and becoming experts of that application or website by experience.

As a consequence of the pendulum swinging back and forth, there are now debates surrounding the use of novices and experts in usability studies (Kjeldskov et al., 2010), often dealing with questions as to when and where it is most appropriate to involve what type of users and with what level of expertise. It is agreed upon, that there is a difference in test results, based on the expertise of the users used in tests and experiments. In *Does Time Heal? A Longitudinal Study of Usability*, Kjeldskov et al. (2010) tested how the effects of usability factors are changed over time, when conducting the same usability evaluation with a group consisting of novice users in the medical industry. They were evaluating usability in a newly implemented electronic system for administrating hospital patient records, and the purpose of their paper was to inquire into the difference in how novice and expert users evaluate usability differently, by testing the same group of users in the same system and with the same parameters, but 15 months apart. The paper went with the same principle as with Nielsen's definition, where the difference between a *novice* and an *expert* is found primarily within the technical expertise of that person, rather than basing on the domain knowledge. The nurses that were chosen for this evaluation were all trained nurses (and therefore with some level of domain expertise), but they were novices in the sense of using the newly implemented medical records system they were evaluating.

The results of the paper are interesting. In terms of effectiveness on the system, after 15 months, the experts solved significantly more tasks on the system, than they did as novices. There was also less

[6] Quick examples are Google's Calendar, Microsoft's Office 365 or Google Drive which are all online applications that does not require any software installed, and can be used on any platform.

variation in types of tasks being solved among the experts (Kjeldskov et al., 2010, pp. 5-6). In terms of usability, a total of 103 usability problems were identified by the authors (usability experts) beforehand. Of these 103 problems, novice users experienced 83 of them, where the expert users only experienced 63 in total. The novices also identified all critical non-unique problems. They conclude that: "*The implications of this finding are debatable. On one hand it can be stated that one should use novices because they enabled more problems to be identified. On the other hand, it could be argued that the use of experts supported the elimination of noise from "false" usability problems. Regardless of which of these points of views one may subscribe to, however, our results show that when evaluating a system designed for highly specialized domain, including users who are novices with the system but highly experienced with the use domain as test subjects can support the identification of as many critical and serious usability problems as when using system experts. This finding is important in situations where expert users may be a scarce or non-existing resource*." (Kjeldskov et al., 2010, p. 8).

Similar results are concluded in an exploratory study by Karapanos et al. (2008) where they measured how users form evaluative judgments during their first experience with a product (they used a Smart TV remote in the study) and how these judgments were different after owning and using the product for four weeks. The study showed that during their first experience with the product, the test participants rated the product on *beauty* and *goodness* the most (which include variables such as *practical*, *manageable*, *presentable*, *innovative* and *simple*). However, after the four weeks, their ratings had changed perspective and now focused primarily on how the product was usable for them in their everyday lives (or how it was not). "*Eventually, users were not any more surprised by the product's stimulating character and the product's novelty lost its power to make the product more beautiful in the users' eyes*." (Karapanos et al., 2008).

In another study, examining the effects of display blurring during program debugging on novices and experts, Bednarik and Tukiainen (2005) showed that blurring certain parts of a computer screen used for programming (the parts that were not focused on, based on real time eye-tracking technology), experts were found to find the blurring more of an obstacle than novices, indicating that experts are able to process more information with their peripheral vision. For the novices in the test, the screen blurring did not matter as much, since they were already only focusing on one thing at a time (Bednarik & Tukiainen, 2005).

In 2000, Nielsen ends his article with summing up that there are still valid reasons for continuing to support novice users on website usability design, but that he predicts that the pendulum "*will soon start swinging a little bit in the other direction, even if it won't swing all the way back to a single-minded focus on experts*." (Nielsen, 2000). Nielsen's prediction seem to have been absorbed into much usability evaluation and Information Architecture literature, indicating the importance of including users that match the target group profile defined for the system or website being developed on.

It is often emphasised how important it is to make sure that the group of participants used for evaluations are part of a carefully selected group of users that match certain criteria. The criteria should be based on the product being evaluated, and should be as specific as possible in order to reduce the chance of utilising participants that do not match the target group profile (Tullis & Albert, 2013, pp. 58-59). Rubin and Chisnell (2008) even claim that test results will only be valid if the participants are representative of the intended users' background and abilities (Rubin & Chisnell, 2008, p. 115). They do, however, also recognise that the definition of novice and expert users (domain or technical) can be somewhat ambiguous if not defined clearly from the beginning, and that it is a good way to classify users and their experience with a certain product or service. They also advise that the classification of users as novices or experts should be used internally in the design group, rather than for recruiting participants, as people might often have a different opinion of their expertise level, than the level of expertise criteria defined by the design team (Rubin & Chisnell, 2008, pp. 119-121).

In other types of tests, for example heuristic evaluations that are conducted by evaluators, and not actual end users or participants, there has also been shown a big difference in the results of those evaluations, based on the usability expertise of the evaluators. Quantitative results showed that the expert evaluators found many more usability problems (both minor and major), and also found them much quicker (Botella et al., 2014; Dou et al., 2009; Kinney et al., 2008; Nielsen & Molich, 1990a; Nielsen, 1992). In their study, Dou et al. also compared the difference between two groups of users with different expertise levels, who were used for an evaluation of a visual analytics tool for investigating financial fraud. None of the participants had any experience using the tool, but had either many years of experience within the financial sector, or were graduate students with limited experience yet, making them either domain experts or novices.

Much like the earlier examples, the results of the study showed that the financial experts found it much easier to begin using the application, as they found it less frightening to utilise their domain expertise to express eagerness and curiosity to explore the application and its functionalities. The novices were less courageous, and were more prone to ask the observers of the Think-Aloud tests to help them proceed (Dou et al., 2009).

Understanding the mind-set of why it is important to include participants that are, in one way or another, experts within the context of evaluation is not hard. Clearly, the results of evaluations, whether they include users as test participants or are purely heuristic and based on evaluator expertise, are greatly influenced by the level of domain expertise within the context areas of what is being evaluated. This is the reason for why it is often stressed how important it is to recruit participants that fit into the target group of what is being evaluated, and that it is equally important that the participants fit criteria that were predefined (Sova & Nielsen, 2003, pp. 29-31). The unanimity is that domain expertise in most instances enhances performance of users, compared to those with little or no domain expertise. It is much more problematic to find information about why the novice users in many cases are being overlooked still. The agreement seems to be, that if the users are not experts within the field of evaluation, or fit the criteria of ideal participants, they are not as useful, or useful at all, for evaluations (Jenkins et al., 2003; Rubin & Chisnell, 2008, p. 115).

Even though this seems to be the case, it is interesting to notice that there are studies showing the contrary too. In *Differences in Novice and Experienced Users in Searching Information on the World Wide Web* (2000), Lazonder et al. recognise that "*Research consistently shows that domain expertise enhances search performance. Without exception, studies report superior performance of domain experts over domain novices in terms of efficiency and effectiveness.* […] *That is, experts take less time to complete the search tasks and produce a greater number of correct solutions*." (Lazonder et al., 2000). In their own research, comparing information search processes of technical experts and novices (experience using the Internet being the context of expertise) using the Internet, they concluded that *domain* knowledge (the context of the information they were asked to search for) was not the most significant factor, but that the technical novices and experts did equally well when asked to locate information on a specific website (amount of time and number of actions needed). The technical experts did however locate the specific website faster than the technical novices (Lazonder et al., 2000). These results are interesting for our study, as it deals with user groups that are not differentiated by their technical expertise, but rather their domain knowledge within a specific field.

2.8 Summary of Theoretic Background

As a term, usability can be hard to define precisely. It is dependent on the product it is being applied to, and the term's different attributes are not all equally relevant in any case or example. For this reason, when trying to measure or evaluate the usability of a product, it is important to thoroughly consider which methods can yield the most useful dataset on the attributes that you want tested. This also includes considerations on the best setting for conducting the usability tests, depending on the focus of the test and which methods were selected. If the focus is on pure textbook-usability, a controlled laboratory setting can help eliminate outside influences and disrupting, and allow the test participants to complete a very consistent testing process. If the focus is more on the product as a whole, and you want to detect the advantages and disadvantages of using it in the wild, or in real life situations, a natural setting should be considered. This setting would yield better results on how the use environment influences how end users might use the product, but is easier disrupted by things that you cannot control, and the test data can be harder to document and analyse.

Another very important factor is to classify and differentiate the target group(s) of the product that you want usability tested, as this is very relevant when recruiting participants for the tests later in the process. One of the popular ways to do this in usability studies, is to classify them by their domain and technical expertise, depending on the product being tested and which of the two are important for the study. Numerous studies conclude, that it is important to use test participants who have expertise levels on the relevant attributes, but there is still much ambiguity on when and if novice participants are good choices for usability evaluations. There is some indication showing that there might be areas of usability testing where novices can yield better and more plentiful data than experts. However, this aspect seems to have been deemed non-essential in many cases, and the prevalent opinion is that you should only be considering actual end users, or users with a high level of relevant domain expertise as participants for usability testing.

Our hypothesis revolve around this subject, and the evaluation of Feriecenter Slettestrand's websites is an excellent context for testing our assumptions on the importance of including novice participants in usability testing to see in which areas they contribute something that you might not have obtained or expected to see with expert participants only, and how participant domain expertise can influence the data collected during usability and/or Information Architecture evaluations. In the next section, we will go into details about the methods that we have selected for our usability testing process, and also enlighten on our participant recruitment process and evaluation research design.

3.0 Methods: Introduction

To conduct an evaluation of a given system, the owners of the given system and the usability experts who are going to conduct the evaluation, have to decide what the goals of the evaluation will be and what questions the goals will answer. When doing a usability evaluation, these questions will be verbalised and made clear before conducting the evaluation (Rogers et al., 2011, p. 456). The questions will help determine what usability methods the evaluators should use, which also puts up a series of practical issues that needs to be addressed (location, technology, equipment etc.) along with requirement confronting ethical reflections, if the evaluation process involves test participants. The process of analysing, interpreting and presenting the data acquired from the methods, requires both time and expertise, as to give a meaningful evaluation result, which the owners of the given system can benefit from. To make sure that these considerations are taken into consideration and used to make a well-structured approach to the evaluation, we have chosen the DECIDE framework (Rogers et al., 2011, pp. 456-475) as our guideline. The framework consists of six steps, which can be followed in a non-consecutively/iteratively order, as each section influences the others. The six sections are as follows:

1. Determine the goals
2. Explore the questions
3. Choose the evaluation method
4. Identify the practical issues
5. Decide how to deal with the ethical issues
6. Evaluate, analyse, interpret and present the data

The goal (Rogers et al., 2011, p. 457) for our evaluation of Feriecenter Slettestrand's main website and mountain biking website is divided into two parts and have been discussed earlier; the first part is an evaluation of Feriecenter Slettestrand's websites, where we will end up with data, derived from our test results, to how the usability and Information Architecture of their websites can be improved. The second part is the analysis of these test results, which will give us an insight into how domain expertise affects evaluation test results. The evaluation of the websites is secondary for us and not explicitly relevant, but the data that we derive from it, will be the foundation of our analysis.

This leads to the selecting of evaluation methods. The test methods that we will be testing our hypothesis with are Think-Aloud and Card Sorting. The Think-Aloud method is relevant for our hypothesis, as it is one of the most popular test methods, when conducting usability evaluations with

users on websites (Nielsen, 2012; Snitker, 2001, p. 95; Tullis & Albert, 2013, p. 102) and has been so since the early nineties (Nielsen, 1993, p. 195).

The Think-Aloud method used for testing interfaces, has its beginning in the early eighties (Lewis, 1982), but the method of "thinking aloud", is mentioned as far back as the nineteen twenties (Watson, 1920). Danish companies like SnitkerGroup[7] and Usertribe[8] utilises Think-Aloud when testing websites and systems containing an interface, where usability is highly regarded. Think-Aloud as a strategy is also employed by the researchers we mention in our theory section referring to other researchers, trying to explore the difference between experts and novices (Bednarik & Tukiainen, 2005; Dou et al., 2009; Jenkins et al., 2003; Kjeldskov et al., 2010). Think-Aloud testing is furthermore a go-to usability test method for professional usability labs, as seen in a comparative usability study with nine independent usability labs, conducted by Molich et al. (2004). The usability labs participating in the study were free to choose their own preferred method for the usability study and eight of the nine usability labs chose a variation of the Think-Aloud method. This all contributes to the idea of us using the Think-Aloud test method, when testing our hypothesis, as it is a highly valued method in the professional world of usability testing. This makes it interesting to us, as we can contribute to the theory of this method, by demonstrating how the test results could be affected, when using the method with domain novices, and how the results compare to the results of domain experts.

The Think-Aloud method is useful in many circumstances and can be incorporated into various testing scenarios and not only Information Architecture relevant scenarios, but as we also want to test our hypothesis on a specific Information Architecture test method, we have chosen to use the Card Sorting test method, as it is one of the most valuable test methods in Information Architecture (Morville & Rosenfeld, 2006, pp. 255-259). This method has also been used for a long period and has been an acknowledged test method since the eighties (Tullis & Albert, 2013, p. 218). The Card Sorting method also relies on the test participants and their ability to use their experience (mental models) and their ability to think and organise (Morville & Rosenfeld, 2006, pp. 255-259; Petrie et al., 2011; Rubin & Chisnell, 2008, p. 18).

However, the DECIDE framework should not be considered a vital part of this thesis, but rather just a basic framework that shows how we have decided to work with and structure this thesis.

[7] http://snitkergroup.com
[8] http://usertribe.dk/

3.1 Heuristic Evaluation

Before conducting the Think-Aloud and Card Sorting methods, we need to understand Feriecenter Slettestrand's two websites in regard to navigation, labelling, labels, interactivity, content etc. To do so we are going to use the method Heuristic Evaluation (Nielsen & Molich, 1990a; Petrie & Power, 2012). This will help us when making the research design for the Think-Aloud and Card Sorting tests, as we can focus those tests on elements of the websites that the Heuristic Evaluation can point out to be advantages and disadvantages.

To complete a Heuristic Evaluation, you need to find usability experts, who can be assigned the roles of evaluators. Those evaluators are going to use their expert knowledge to analyse and report on the given website they are presented with. To ensure that the evaluators examine the website in an equal manner, they are assigned with a guideline (Nielsen & Molich, 1990a; Petrie & Power, 2012). Nielsen and Molich (1990a) developed one of the first guidelines for evaluating systems. Their guideline contains a list of nine heuristics, which the evaluators are to follow to ensure that they cover every aspect of the given system. These nine heuristics have since then been the basis of future updated versions of the Heuristic Evaluation method, as it has been adopted by other researchers (Budd, 2007; Instone, 1997; Petrie & Power, 2012). Nielsen and Molich's nine heuristics were compiled from research on interfaces and not websites in particular (Petrie & Power, 2012), which makes their heuristics less relevant for us. We have therefore chosen Petrie and Powers' (2012) Heuristic Evaluation guideline, which consists of 21 heuristics. It is interesting for us to use, as it is one of the most recent completed Heuristic Evaluation guidelines (which ensures its relevance) and as it also has a focus on Information Architecture, which is relevant for us in regard to both the Think-Aloud method and in particular the Card Sorting method. Petrie and Power's heuristics are a combination of four overall categories that focus on *physical presentation*, *content*, *interactivity* and *information architecture* (Petrie & Power, 2012). The practical design of how we will use the Heuristic Evaluation will be presented in the Research Design section.

3.2 Think-Aloud

As mentioned earlier the Think-Aloud method is the most used method for usability testing (Nielsen, 1993, pp. 195-198). Think-Aloud involves letting test participants think out loud while performing tasks given by a facilitator, while operating on a given interactive interface (Lewis, 1982), for example a website. The exercise of thinking aloud was first used as a psychological method (Ericsson & Simon, 1993, pp. 1-10), but in the beginning of the 1990s it was being used to evaluate interactive interfaces, when Denning et al. (1990) used the method for testing systems developed by Microsoft. The method can be relatively cheap to conduct and it can result in a large amount of qualitative data from a small number of participants. The data includes statements and thoughts by the participants, which can be used as a strong argument for the (re)design of a website (Nielsen, 1993, pp. 195-198).

3.3 Procedure for Think-Aloud Tests

Think-Aloud tests are usually conducting in a laboratory setting, where the participants complete a set of predefined tasks, while thinking aloud. When the participants are thinking aloud the facilitator will have to closely watch, listen and note what the participants are doing (Kalbach, 2007, p. 162). Some usability experts mention that the facilitator will have to be able to see if the participants are struggling with a task and then be able to guide them, if they are completely stuck on a specific task (Rubin & Chisnell, 2008, p. 55). The tricky part is then to know when to interrupt the participants, as this is a critical point in the test, the facilitator will have to be careful not to interrupt to early as a silent participant can either be because the participant is just thinking about something, or that he is stuck. The findings that occur when the participants are struggling are crucial and important issues to handle later on when redesigning the system (Rubin & Chisnell, 2008, p. 55). The facilitator can also choose to conduct Think-Aloud test, without having to observe and interrupt the participants, but instead let them finish their tasks alone. This is done by giving the participants a given amount of minutes for each task (Kalbach, 2007, p. 162).

The number of participants depends on the usability expert you console to. Rubin and Chisnell argue that four to five participants are enough to expose the vast majority of usability issues (Rubin & Chisnell, 2008, p. 126). When exceeding the five users you will start to see a lot more of the same results than with your first five participants, making the extra effort redundant. This also takes the economic aspect into account, as usability testing can be an expensive and time consuming process (Nielsen & Landauer, 1993). These five participants can then be used to participate in an iterative testing. The number of participants should however be decided with the purpose of the Think-Aloud

method in mind, as there is a difference in testing a small website and a large website (Kalbach, 2007, p. 162).

When the number of participants has been decided, the facilitator will have to design the tasks. The tasks that the facilitator gives the participants can focus on different elements of the information architecture of the website (Kalbach, 2007, p. 162):

- **Visibility**: Do the participants see important elements on the website like the "Sales" menu in a web shop or the "Login" label on a municipality website.
- **Labels**: Do the participants understand the labels that they find on the website or do they expect different content when clicking on a label. For example, does participants expect to see a "Book a room", when clicking the label "View available hotel rooms"?
- **Orientation**: Can the participants orientate themselves at all times, when navigating a given website or do they become lost and express that they do not know where they are on the website. What strategies do they employ to solve this?
- **Findability**: Is the content on the website findable by the participants or do they browse back and forth looking for certain content?
- **Efficiency**: Is it time consuming for the participants to use a given website or can they locate and use content within reasonable time, when solving tasks that explore standard goals of that site?

As mentioned in the theory section, a usability test like Think-Aloud can be conducted in various settings. When conducting Think-Aloud on a website you will however need to use a computer (unless you are in the early phases of development and still at a paper and pen level), to make the participants interact with. This can make the practical part of the method easier, as the interactions (cursor movements, clicking, and typing) and the verbalisation of thoughts the participants express, are recorded via the computer. The need for note taking is therefore reduced.

A number of software programs have been developed for this, but it can easily be done using already installed software on your PC or Mac. On the Mac the video recording and playback program QuickTime has a built in functionality that lets you record both what happens on the screen and the sound of a microphone, which is then stored on your Mac. QuickTime is therefore useful when conducting the Think-Aloud method, as it is easy to use, free, and records high quality videos with microphone audio, which is very useful for later analysis.

There are however still many aspects that needs consideration when conducting a Think-Aloud test session, which are often determined by the purpose of the test. However, even though there is a wide range of approaches, there are some basic steps that are repeated often (Kalbach, 2007, p. 162):

- **Participants**: Find appropriate participants (as mentioned earlier, usability experts recommend using participants that can be categorised as intended end-users).
- **Protocol**: Create a plan for the Think-Aloud session, which includes the actual protocol of the tasks that the participants are being instructed to follow.
- **Setting**: Establish a laboratory (this can be interpreted in various ways) to use for the conducting of the Think-Aloud session. This includes the practicalities and instrumental setup (Computers, microphone, recording software (QuickTime), coffee and gummy bears (for the participants, *of course*), pen and paper for note taking, testing of the setup, etc.)
- **Test**: Introduce your participants to the Think-Aloud method. This introduction can include the facilitator(s) showing how the test is going to proceed, by showing what they expect from the participants. This can be done by making a two minute Think-Aloud test on a random website. When the participants are ready, you conduct the test with your participants (one at a time), following the predefined protocol.
- **Analyse**: To understand the results you receive from the test you analyse the recordings and test notes, and through that extract the findings that the participants encountered during the test.
- **Presentation**: The results of the usability testing should at this point complete the goals that were set for the Think-Aloud method and you will be able to present, explain and suggest improvements for the tested website.

3.4 Challenges with Think-Aloud

It is important that an evaluator with some knowledge of the method and usability in general assesses the Think-Aloud data, as the participants may conduct their own opinions on why a system gives them difficulties. The evaluator will then have to interpret the real issues of the system and why the participants struggle when trying to use the system (Nielsen, 1993, pp. 195-198). Participants may say that the colours of the website is the reason for them not finding the content they are looking for, when in fact the facilitator can see by their actions that the navigational labels are at fault. Some participants may also find it difficult to think aloud while performing the tasks, as it may not seem natural to them (Nielsen, 1993, pp. 196-198; Rubin & Chisnell, 2008, p. 54). It is however still worth

the trouble, as the verbalisations from the participants (thoughts, actions, reactions, delight, fury etc.) are valuable in understanding the participant's experience with the system (Barnum, 2011, p. 19).

A study by Berry and Broadbent (1990) showed that users may work faster than normal, when thinking aloud, but Rubin and Chisnell mention the contrary; that time measuring should be avoided in Think-Aloud tests, as the participants work *slower* than normal (Rubin & Chisnell, 2008, p. 54). We chose to include time-measuring for measuring the usability in our analysis, as we believe the truth might lie somewhere in between the two, and that participants might work faster because thinking aloud allows them to think more clearly on how they approach the task they are given, but that they may also spend a little extra time explaining the process for us, rather than just keeping it personal.

A study by Wright and Converse (1992) further shows that participants only found 20% of the errors that other participants found while doing the same test silently. Some participants may also stop themselves from saying various things, which can be both a conscious or unconscious choice (Rubin & Chisnell, 2008, p. 54). The participants may also start to verbalise ideas of how the problems they encounter can be dealt with, which can be a positive, if this is what the test is going for, as it can help the evaluator in their work with coming up with new ideas (Rubin & Chisnell, 2008, p. 54).

3.5 Concurrent and Retrospective Think-Aloud

The Think-Aloud method described above is the *concurrent* type (Ericsson & Simon, 1993, p. 16; Tullis & Albert, 2013, pp. 81-82), which means that the participants think aloud while interacting with the given system. This method of conducting a Think-Aloud test is the most relied on by usability researchers (Tullis & Albert, 2013, pp. 81-82).

Another version of the Think-Aloud method is the retrospective review method (Ericsson & Simon, 1993, pp. 19-20; Rubin & Chisnell, 2008, pp. 54-55), which is a method that tries to deal with some of the issues mentioned above with the Think-Aloud method. It does this by letting the participants work through the task that they are given by the facilitator without having to think aloud while doing so. The facilitators' job is then to note every time the participants have issues with the system being tested. While the participants are working through the tasks, the facilitator will record the scenario. The recordings will then be used after each of the participants have finished their tasks. The facilitator will then go through the recording with each of the participants and stop when they encounter the issues that the participants met in their work with the tasks. The participant will then be asked questions, which allow for them to explain what their issues were at that given point and how they

work around those issues. Rubin and Chisnell encourage retrospective review when using participants who can have difficulties in verbalising while conducting the tasks, such as children, people with disabilities or elderly people (Rubin & Chisnell, 2008, pp. 54-55). The retrospective Think-Aloud method is however gaining popularity as for example Petrie and Precious (2010) found that a retrospective Think-Aloud method gave more emotional responses, than a concurrent Think-Aloud method. So for measuring the participant's emotions during usability studies the retrospective method might be of more use than the concurrent. The retrospective method does however have two serious disadvantages; firstly, the method is more time consuming than the regular concurrent Think-Aloud method (Tullis & Albert, 2013, pp. 81-82). Secondly, it can be difficult for the participants to recall what they were thinking when re-watching themselves solving the tasks, as the extra time from when they were using the system till they are interviewed by the facilitator, brings the risk of participants forgetting their initial thought process or rationalising their behaviour and therefore not report what actually happened. The facilitator will then not be able to extract the participant's thoughts from when they were trying to solve the problem, and the data will be lost (Elling et al., 2011; Rubin & Chisnell, 2008, p. 55).

This can however be avoided (in some cases) by using eye tracking technology (Freeman, 2011; Guan et al., 2006; Olsen et al., 2010). The technology allows for the facilitator of the Think-Aloud test to observe where the participant is looking when completing the given tasks. However, the study by Elling et al. (2011) shows that there are no difference in the use of eye tracking and reports that the eye tracking apparatus may even distract the participants from the actual task at hand, obscuring the results given by the Think-Aloud method.

The combination of a retrospective Think-Aloud method and eye-tracking is interesting, but we will focus on the concurrent Think-Aloud method alone, as the research we have gone through, which investigated the differences between novices and experts, have not used neither retrospective Think-Aloud nor eye-tracking technology. Jenkins et al. (2003) and Dou et al. (2009) do however interview their participants retrospectively after the concurrent Think-Aloud testing of the given systems, which is a technique that we will also use, especially for measuring user satisfaction as a part of the overall usability.

3.6 Card Sorting

The reason for conducting the Heuristic Evaluation was, as mentioned earlier, to get information on what parts of Feriecenter Slettestrand's main website and mountain biking website that are well functioning and which parts that can lead to issues for the users. The information from the Heuristic Evaluation will help us in designing our approach for the Think-Aloud method, as we can target the usability problems that we encounter in the Heuristic Evaluation. This Heuristic Evaluation will also create a foundation for our approach to the Card Sorting method, as the navigation, labels and content, which are the main focus point in Card Sorting (US Department of Health and Human Services, 2015a) are significant in the Heuristic Evaluation by Petrie and Power (2012).

As well as Think-Aloud being the most popular usability test method, Card Sorting is one of the best user testing methods used in Information Architecture, for learning the mental models of the participants, which shows how they sort, label and group content (Morville & Rosenfeld, 2006, pp. 255-259; Petrie et al., 2011; Rubin & Chisnell, 2008, p. 18). What makes Card Sorting one of the best and most used methods in Information Architecture contexts, is especially its ability to envisage how to organise and design a website or interactive interface and at the same time it is because of how easily it can be conducted (Maurer & Warfel, 2004; Spencer, 2009, p. 10).

First, you will have to write the labels from the content and main-and submenus on the cards along with a number (used to make the analysis easier). The cards are then given to the test participant, who is asked to group them in a way that fits the purpose of the research (Morville & Rosenfeld, 2006, pp. 255-259). The purpose can be to build a structure of the website, define what should be on the front page and examine what labels to use for the navigation (US Department of Health and Human Services, 2015a). When designing a navigation system for a website that sells clothes, you could then for example tell the participants to write down labels that they expect to find on such a website. They will then be told to sort the cards in ways they find meaningful and when all the participants have done so, you will then compare the results and develop a navigation that fits those participants' mental models.

There are various ways of conducting a Card Sorting and to choose which is relevant for your research depends on the research area of your interface and the outcome you are looking for. First, you will need to decide on which way the participants shall order the cards. This can be done in an almost endless ways, whereas Morville and Rosenfeld suggest the following methods: open/closed, phrasing, granularity, heterogeneity, cross-listing, randomness and qualitative/quantitative (Morville &

Rosenfeld, 2006, pp. 255-259), where the two most used methods are the open and closed methods (Tullis & Albert, 2013, p. 218). Open Card Sorting lets the participants make their own labels and closed Card Sorting lets the participant sort cards that a predefined by the facilitator. The facilitator will also predefine the name of the groups in the closed sorting (Morville & Rosenfeld, 2006, p. 256; Tullis & Albert, 2013, pp. 218-219). There are however also hybrids that allows for the facilitator to choose elements from the different variations, for example if the participants have a hard time trying to categorise the labels under one of the predefined categories, then the participants are allowed to create their own categories.

Second, when you have decided on the method for letting the participants order the cards, for example a closed or open-ended Card Sorting, the next step is to decide on which technique you are going to conduct the Card Sorting with. This can be done in a one-on-one session, where the participants sort the cards individually one at a time, while they think aloud and explain the sorting they have chosen. If done in a usability laboratory or in another controlled setting, the facilitator of the Card Sorting will be able to ask questions concerning the way that the participant sorted the cards (US Department of Health and Human Services, 2015a).

The Card Sorting can also be conducted with individuals in a group setting, where the participants are briefed and debriefed together, but work alone. This does however require multiple sets of the same cards or multiple computers if done in a program or through an online Card Sorting Service. The facilitator will however not be able to interact with the participants, when doing it this way (US Department of Health and Human Services, 2015a). The Card Sorting can also be done in a group setting, where the participants talk to each other and through that sort the cards. This can be an effective way of creating main content groupings, but the result can also be conflicted by group dynamics (US Department of Health and Human Services, 2015a). However, the group setting can also be positive, as group discussions can give more insight by participants contributing to each other's ideas (Spencer, 2009, pp. 86-87).

The last Card Sorting technique is the remote or online version of Card Sorting. This technique is becoming more and more popular (Petrie et al., 2011) and there are various solutions for online Card Sorting for example Optimal Workshop[9], Userzoom[10], SimpleCardSort[11] and UXSort[12] to mention a

[9] http://www.optimalworkshop.com/optimalsort
[10] http://www.userzoom.com/card-sorting/
[11] http://www.simplecardsort.com/
[12] http://www.uxsort.com/

few. In the beginning of Card Sorting only *oncard* card sortings were conducted, as Petrie et al. (2011) describe them, but with the development of the Internet, online Card Sorting became more a more popular. The question is however, if the results from the oncard sortings and the online card sortings are the same or if the online version of Card Sorting leads to other results, which conflicts with the reliability of the Card Sorting method. Studies into this problem has been conducted and results from Bussolon et al. (2006), Harper et al. (2002) and Petrie et al. (2011) shows that the difference between oncard and online Card Sorting is insignificant, meaning that the two ways of conducting a Card Sorting are equally reliable and useful to get the same results.

When conducting Card Sortings, a number of elements come to play, as the number of groups, the accuracy and the labels created by the participants' play a huge role and is the main concern of conducting the Card Sorting, either oncard or online. This is however not an issue as Harper et al. (2002) discovered in their research on the topic, where they compared the results of a live Card Sorting and a software version called Team Performance Laboratory Knowledge Analysis Test Suite (Harper et al., 2002). The reason for online versions or installable software to be used on conducing Card Sorting is because of the amount of work put into an oncard Card Sorting, where there is a common agreement that the work of doing an oncard version of Card Sorting can be tedious and time consuming (Harper et al., 2002; Petrie et al., 2011).

There are however differences between the two methods for the facilitators and the users, as the study by Harper et al. (2002) showed that first-time participants of online Card Sorting used more time sorting the cards, as compared to the oncard version. If the participants are experienced with the online version they will use less time finishing the online Card Sorting. The time spent by facilitators is reduced significantly when using the online version, as the work preparing for an oncard Card Sorting takes a considerable amount of time (Petrie et al., 2011).

As a facilitator you will then have to consider the time issue, for example how much time you will use on creating the Card Sorting, where your participants are located and how experienced they are with Card Sorting. The time that you will need to take from the participants' own time and the time you have to spend on creating the card sorting is therefore important to consider. If the participants are first time users and you will not be using those participants later on in your research, it will be reasonable to conduct an oncard Card Sorting.

The experience we have had using Card Sorting have left us to prefer the oncard version, where we are able to conduct the Card Sorting and as facilitators able to ask the participants about their choices.

We have used an oncard Card Sorting in our 8th semester project, where we were designing a new website for a kindergarten. We used a closed Card Sorting with 10 participants. The results were used to decide on the labels for the navigation and how the navigation should be categorised. During the same semester, we used online Card Sorting for a class assignment and found it to be a less exciting method.

When a fitting method of Card Sorting has been chosen, the next step is to choose how many participants are appropriate for your evaluation. In our past experience, we conducted tests with 10 participants, but how many participants is the "correct" amount? The number of people who are going to participate is up to the researchers, but to make the most from a Card Sorting, Nielsen (2004) recommends using about 15 participants. The studies from Nielsen Norman Group[13] show that 15 participants lead to a satisfying level of correspondence. The reason for the number of participants being three times bigger than the normal recommendation of four to five participants (Rubin & Chisnell, 2008, p. 126) is that Card Sorting is a *generative* test method compared to an *evaluation* method like Think-Aloud. Card Sorting does not let the participants look at a design and make them say "this works", "this does not work" or "I don't understand this". Instead, the goal is to understand how the participants think. The mental models and vocabulary of the participants are very different, leading to more complex results. In order to get results that can lead to a conclusion, you will then need more than the standard amount of five participants (Nielsen, 2004). The best practice when conducting Card Sorting is to use 30 to 40 cards, as not to overwhelm the participants with cards (US Department of Health and Human Services, 2015a).

We are going to conduct a hybrid between a closed and an open Card Sorting, as we believe this to be the most relevant Card Sorting for our case. The cards will have labels from Feriecenter Slettestrand's website on them, and the participants will be given the task of sorting them into groups. We will not give them fixed categories as in a regular closed Card Sorting. Instead we are going to let them create as many groups as they see fit and then ask them to choose one card from each of their groups as the main category for that group. We do however know from Information Architecture theory that when creating a navigation menu system, the best practice is to implement a maximum of five to seven categories in the main menu (Alves, 2011).

[13] http://www.nngroup.com/

3.7 Card Sorting data

When conducting a Card Sorting the data that the method yields needs to be analysed as to understand the sortings that the participants have made. The method for this is different when conducting a closed or open Card Sorting, so you will therefore need to use an analysing method that fits the variation of Card Sorting that has been used.

In a closed Card Sorting the results you will look for, is how the users sorted the given cards into your predefined categories. What you want to go for is a percentage of 100 for each of the cards in one of the groups. This is however not very likely to happen. The ideal method of choosing which group that will give the best result, is to look for the group that is "winning" (Tullis & Albert, 2013, p. 225). As seen in the example below, card number three almost has a percentage of 100 and that is the percentage you are hoping for. Card number two however will be difficult, as it is almost a tie between Group A and Group C.

Card number	Group A	Group B	Group C	Max Percentage
Card number 1	12%	75%	13%	85%
Card number 2	48%	3%	49%	49%
Card number 3	2%	5%	93%	93%

This is however a tedious method for presenting Card Sorting data, as it is slow and does not go into detail with the various aspects of the how the cards have been sorted. We want to get as much information as we can out of the sortings the participants make, to make sure we can verify our results, for example the differences and similarities between the data from domain novice and expert participants.

We have therefore been looking for further ways of presenting this data: The data can then be further analysed by conducting a standard statistically method for studying similarities. For this purpose we have acquired a spreadsheet[14] for Excel by Spencer (2009, pp. 175-103). The spreadsheet is free to use, and is usable for Card Sortings with up to 40 participants and 200 cards. The spreadsheet is

[14] http://rosenfeldmedia.com/books/card-sorting/

therefore relevant for us to use, as we are using less than 40 participants and less than 200 cards, as to stay within the recommended numbers by Nielsen (2004).

Card no	Card name	AKTIVITETER	BOOKING	FAMILIEN KRONBORG	HANDICAPVENLIGT OPHOLD	KURSUS CENTER	KURSUS CENTRETS FACILITETER	OM OS	OPHOLD	PRAKTISK INFORMATION	PRISER	RESTAURANT	VI TILBYDER	VÆRDAT VIDE
12	Specialkost									10%		90%		
13	Booking		20%			10%		10%	10%	20%	10%		20%	
14	Vin og gourmetaften											100%		
15	Aktiviteter	40%				10%			10%				40%	
16	Familien Kronborg			10%		10%		70%						10%
17	Find Vej					10%		50%		20%			10%	10%
18	Julefrokoster											90%	10%	
19	Aktiviteter i naturen	40%					10%		10%				40%	
20	Ferieboliger		10%			10%		10%	10%	20%			40%	
21	Indkøbsmuligheder					10%		20%			10%	30%	20%	10%
22	Handicapvenligt ophold		11%		33%		11%			11%			33%	
23	Åbningstider restaurant		10%			10%				10%		70%		
24	Teambuilding	20%				10%	20%						50%	
25	Boenheden Delfinen		10%	10%	20%	10%	10%	10%		10%			20%	

Figure 1 - An example of how the percentages of agreement are visualized in the spreadsheet by Spencer (2009, pp. 175-203)

From here on you start to analyse the groups that the participants have created, as to see which groups that were formed, if there were big differences in the content of the groups and if your expectations of the Card Sorting were met (Spencer, 2009, p. 188). When looking at results from a closed Card Sorting the results can differ a lot from what you were expecting of the Card Sorting. Furthermore if a large amount of the cards are sorted into a few of the predefined categories in a closed Card Sorting, it can be an indication of those card labels being too broad. If labels have been placed in many of the categories, those categories' content may not be clearly defined or it may contain similar content (Spencer, 2009, pp. 201-202). Spencer's Card Sorting spreadsheet is also useful for keeping track of the data from the sortings (so it is not necessary to look at a bunch of photographs of cards), includes the number of unique and total cards in each category and how many participants who used each category. It also keeps track of all the categories used in the sortings, both the standardised categories that existing card labels were used for, but also the categories that the participants will have a chance to create themselves. In the research design section 4.8, when we present the data from the Card Sortings, we will elaborate on this spreadsheet, its use and how we can derive empirical data from it.

3.8 Domain Knowledge and Feriecenter Slettestrand

For this thesis, we use two classifications of user groups, primarily defined by their level of domain knowledge within mountain biking. As described earlier, one of Feriecenter Slettestrand's primary target groups are mountain bikers who visit the resort to ride the nearby mountain biking tracks. The target group consisting of mountain bikers is one of the largest, only surpassed by the group consisting of families or groups that visit the resort strictly for recreational purposes[15]. From one of our meetings with Feriecenter Slettestrand we also learned that they are afraid of the mountain biking target group being the dominant, and losing traffic on their main website because of it, indicating that the mountain bikers take up a significant part of the overall group of users of their websites.

3.9 Target Group: Mountain Bikers

As already mentioned, one of the target groups that we have put most of our focus with in this thesis, is the group of mountain bikers who visit Feriecenter Slettestrand's websites primarily for activities related to mountain biking. This group is characterized by their deeper understanding of mountain biking as a domain, making them the group of test participants with more domain expertise.

Microsoft Research have conducted a study, exploring how to define domain expertise based on web search behaviour (White et al., 2009), and concluded that one of the ways to define domain expertise is to understand in which ways they are able to use domain specific terms and queries in information search sessions on websites or services that are primarily aimed towards their domain of expertise. Even though we realise that mountain biking as a domain might not be too abstract for people with little domain knowledge to understand most of, there are still content elements and options for queries that contain domain specific terms available in the domain. These specific terms and queries might not be used by people with no or little domain knowledge of mountain biking. For the same reason, users with domain expertise will also have a higher success rate, when it comes to finding information, as they can better relate, validate and interpret information that is domain dependent, than their non-expert counterpart (Kinney et al., 2008; Russell-Rose & Tate, 2013, pp. 26-27; White et al., 2009).

This group of users are chosen as test participants as they are in compliance with the presented literature on how to recruit test participants (Jenkins et al., 2003; Rubin & Chisnell, 2008; Snitker, 2001, pp. 39-40; Sova & Nielsen, 2003, p. 20). They are within one of the target groups of Feriecenter

[15] Based on Google Analytics data of ~530,000 sessions of unique visitors. Of these sessions there have been ~84,000 unique visits on the main page, ~29,000 unique visits on the page dedicated to information on the holiday homes and ~33,000 unique visits on the page dedicated to mountain biking (multiple visits per page for each session are not included).

Slettestrand's websites, have a strong domain knowledge within a specific domain that is used on the websites, and they fit the criteria of a representative user group. We will later introduce and specify the evaluation participants, and the participant criteria, in more detail.

3.10 Target Group: Recreational Guests

As exemplified with the study from Lazonder et al. (2000), domain knowledge, or domain expertise, might not always be the most important factor, when evaluating websites – especially not on websites, like Feriecenter Slettestrand's, where there are several target groups and domains to take into account. To test our hypothesis on how usability and Information Architecture evaluation results differs between groups of users with different levels of domain expertise, and where the results are applicable, it is necessary for us to include a group of participants with no or little domain expertise. This group is a more *generic* or *all round* group of users, as it is much harder to classify. The reason for this being that the target group of the recreational guests on Feriecenter Slettestrand's websites is a much broader target group. It includes families, couples (with or without children) etc., and it is also harder to define an age group, than with the other target group consisting of mountain bikers[16]. This group is not classified by any specific domain knowledge, but will be used as participants in the same tests as the mountain biker target group.

3.11 The Double-Edged Sword

All the participants will complete the same set of tasks in the test they participate in, in order for us to be able to compare and analyse the results. Since we are evaluating Feriecenter Slettestrand's websites, it is also possible for us to use the different target groups to test several subparts of the websites, which might have different target groups, depending on who those subparts were designed for. For example, the mountain biking website is primarily targeted at the target group consisting of mountain bikers, but the main website (which evolves around basic information about the Feriecenter Slettestrand, and all the different services they provide) has a much broader target audience, as it not specific to one topic. These topics include information on staying at the resort, information on the different holiday homes, how the resort is disabled-friendly, information on their restaurant and training centre, what activities goes on etc. Most of this type of information is relevant to every target group considering to use or visit Feriecenter Slettestrand, regardless of reason. It is as relevant for

[16] It is easier to reason that mountain bikers' median age would be lower than the median age of the broader target group consisting of families within all age ranges, as mountain biking can be a rather extreme sport, depending on the discipline.

mountain bikers coming to stay a weekend only for mountain biking, as it is to a family with children with no interest in mountain biking, but are visiting to spend one's holiday recreationally.

Since the mountain biking domain, as described earlier, is not too abstract to understand, even for people without any mountain biking domain knowledge, it still allows the evaluation testing of the mountain biking website with participants from the target group of recreational guests, and helps us substantiate our hypothesis and investigate if that target group is appropriate and practical in usability evaluations like this one. For the same reason, the evaluation case is interesting for us to visualise how and where the different test participant groups differentiate in their evaluation data, as Feriecenter Slettestrand's websites allows for different domains (some are specific, some are not) to be tested by groups of participants with varying domain knowledge.

We expect to see that the mountain biking target group, which is in compliance with prevalent usability evaluation participant criteria, yield useful information and constructive criticism that can be utilised to present valuable proposals for redesign of the websites, but also indicate what about the existing websites works well already. On the other hand, we are also eager to study if the target group consisting of the more generic participants generate data that is still useful for usability evaluations, but also if those participants are able to extract useful evaluation data that might be overlooked by the other group that has more domain knowledge, and if so, why that happens. The assumption is, as exemplified in some of the before mentioned studies (Bednarik & Tukiainen, 2005; Kjeldskov et al., 2010; Lazonder et al., 2000), that the domain novice group might focus more on the aspects of the website that deals with what could be categorised as usability problems (how hard it is to find precise information, comments on layout and design, how information is presented, how efficiency could be improved, subjective satisfaction etc.). However, the domain expert participants might use their domain knowledge to focus on advantage and disadvantage aspects that is more relevant to the target group that they are a part of, and disregard less relevant (more common or generic) issues because of a shifted focus (Petrie & Power, 2012).

In this perspective, the usability testing of the Feriecenter Slettestrand's websites works as a double-edged sword that in one way allows domain knowledgeable participants to express opinions and observations that might be based more on their domain, and less on general usability. In another way it accommodates a group without any specific domain knowledge to do the same, which could give rise to more general comments and observations, and arguably could be more useful for general usability evaluation that serves a larger target audience.

Another reflection is, later in the process when comparing and analysing the data from the user tests, the groups are dependent on each other for comparison, as both groups will complete the same tests, but have different mental models of the domain. Results from the two groups might help validate one another, and confirm or deny our assumptions of the results. We recognise that there is not one single way or method to assess usability and Information Architecture evaluation, and that advantages and disadvantages found in the evaluation are fluid and constantly developing. However, by including several different methods and relevant design expert and user perspectives, we can scope out important and appropriate usability issues.

4.0 Research Design and Data Presentation: Introduction

In the following section, we will introduce and go into details on the usability evaluation process, and argue for the choices we have made during the process. Firstly, we will introduce the criteria for choosing the test participants in our two test groups (the mountain bikers and the more generic group of possible recreational guests). Then we will describe the setting used for the usability tests, and go into detail on the process of each of the methods we have described earlier, and how those methods were used in practice in order for us to complete the usability evaluation of Feriecenter Slettestrand's websites.

Lastly, we will reflect on how the process worked in practice, and reflect on some of the events that we did not anticipate and things that could have done differently if we had the chance to redo the process.

4.1 Recruiting Participants for the Usability Tests

As we have described in the theory section, choosing the right participants is something that is essential to usability testing. To test our hypothesis on how domain novice participants might be useful in usability testing of products targeted at specific domain knowledgeable users, we have chosen to conduct the same usability tests with two different groups of participants; a group of mountain bikers (who has expertise knowledge on the mountain biking domain) and a group of participants with no mountain biking domain expertise.

Our reason for doing so lies with the current idea of always needing expert users, as presented in much literature on usability testing, when conducting evaluations of products with a specified target group (Bednarik & Tukiainen, 2005; Karapanos et al., 2008; Kjeldskov et al., 2010; Nielsen, 2000; Rubin & Chisnell, 2008, p. 115). In many of the studies where the difference between domain novices and experts has been reviewed, they have used the same group of participants twice, but with enough time between the tests, that the novices had become experts with the system they were evaluating. For example, in the study by Kjeldskov et al. (2010), they tested the same group of participants 15 months apart, and in the study by Karapanos et al. (2008) it was just four weeks – most likely based on how much experience with the system or product is required, before the users could be considered domain experts with it. Because of limitations with time, we will conduct our evaluation tests with two groups instead. In this way we can test with already-established domain experts, and users that

are domain novices. We therefore specified the two groups, and they both have the criteria that needed to be filled, before we could use them as test participants:

The mountain biking group needed to consist of participants who have strong domain knowledge of mountain biking. The websites that we are conducting the usability tests on are primarily aimed at two groups, one of which being mountain bikers, the other being a more general group that includes recreational guests, but is harder to define because of the broadness of that target group.[17] The group of mountain bikers meet the requirements that most of the before-mentioned literature agree is the ideal conditions for usability testing, as described in earlier sections: they have a expert level of domain knowledge, and they are representative users who would be likely end users of the websites as well.

The other group does not meet these conditions, as they have been recruited by only one main criteria; they could not have any domain expertise within the domain of mountain biking. They could, however, be in the target group of non-mountain biking, but rather recreational guests (families, couples, etc.). We recruited the domain expert participants through domain specific internet forums[18] and through contacts at Feriecenter Slettestrand. The novice participants however, were a mix of friends, family and a few people recruited through contacts at Feriecenter Slettestrand.

For both groups, it was a given that they fit into a very wide category of possible guests on Feriecenter Slettestrand, and that they had used a computer before, as this was the only technical aspect of the tests we conducted. None of these necessities has been a problem or a big concern in the recruitment process, because of the broadness of these requirements, as we have chosen to only focus on domain expertise. In the study by Kjeldskov et al. (2010), they also included the technical aspect of expertise, as the product they were evaluating in that study was a completely new way of administrating patient journals in a (at that time) newly developed software system used at hospitals and doctor's offices.

As for the age of the participants, we did not specify an age group that we aimed for the participants to fit in, as we were more focused on making sure that they fit one of the participant groups by their domain expertise. We ended up with only using adult participants. The median age for the domain experts ended up being 37,6 years old, while the novices had a median age of 26,4 years. 18 of the

[17] There are other target groups (seminar participants and the disabled) when considering Feriecenter Slettestrand's entire website though, but we have chosen to only focus on the two groups described.
[18] We found several participants on Cycling Nord's forum (http://www.cyclingnord.dk/)

participants were male, and only two were female. The age and sex of specific participants can be found in appendix 05.

The websites we are evaluating for Feriecenter Slettestrand have two target audiences; a rather all-purpose main website meant for all kinds of guests, who would like to know more of any aspect of the resort, and a more separate website targeted at mountain bikers. To be able to compare the data and results from usability tests, which is necessary to confirm or deny our hypothesis, we decided to let every participant go through the same tests in the same order, and with the same testing conditions and requirements.

We have already argued for the optimal number of participants in each of the two methods involving test participants (Think-Aloud and Card Sorting). For the Think-Aloud tests, we ended up using 5 participants from the group consisting of domain knowledgeable mountain bikers, and 5 participants in the group of participants with no relevant domain knowledge, ending up with a total of 10 participants for the Think-Aloud tests. For the Card Sorting, we ended up using a total of 20 participants, where 10 of these participants were the same as the participants used for the Think-Aloud tests, and 10 extra were recruited only for the Card Sorting. This quantity of participants allows us to compare the data of the two user groups with each other, and draw some conclusions based on the data derived from the tests, based on experiences from studies on the subject by Jakob Nielsen and his colleagues (Nielsen, 2004; Nielsen, 2000; Nielsen & Landauer, 1993).

4.2 The Test Plan Sequence

As described earlier, the three methods that we are using are Heuristic Evaluation, Card Sorting and the Think-Aloud methods, and we also completed the tests in that particular order. The reason for completing the methods in that order is, that by executing the Heuristic Evaluation first, we (as evaluators and usability experts) could gain important knowledge on Feriecenter Slettestrand's websites, which would affect the research design of the two usability evaluation methods involving users; We could scope out specific usability possible problems we wanted tested, get to know the existing content, labels and navigation systems.

In continuation of the Heuristic Evaluation, we have also included considerations concerning the websites' existing web traffic, by utilising Google Analytics, to understand how actual end users utilise the existing navigation schemes and labels to traverse the websites, and which elements of the websites that are mostly used and spent time visiting (Clifton, 2012, pp. 54-70). This was done by

looking at the data on unique page-visits per visitor, and comparing how often users visit the different subpages of the websites.

The Card Sorting and Think-Aloud are completed in succession, one participant at a time, but with the Card Sorting done first and the Think-Aloud second, if they were chosen to complete a Think-Aloud test too. We have chosen this order, as the labels on the cards we use for the Card Sorting have been derived from the menu system and navigation labels already existing on the websites that we are evaluating. Because of this reason, we chose to let them complete the Card Sorting first, so that the knowledge of how the websites are already constructed in terms of navigation cannot influence the data collected in the Card Sorting method. This is important, as we wish to utilise the Card Sorting data to evaluate the existing navigation and Information Architecture, based on the participants' current mental models and how they organise information (Tullis & Albert, 2013, p. 48). If we had chosen to complete the Think-Aloud tests, and finish with the Card Sorting, the immediate knowledge that the participants might have gained in the Think-Aloud tests could influence or affect the data in the Card Sorting.

4.3 Heuristic Evaluation

Process and Setting

The usability experts used for the Heuristic Evaluation were the authors of the thesis. The setting was an apartment in Aalborg, which had been transformed into an informal workroom. We sat around a table, where we were facing a 40 inch television, as to make sure that both of us were looking at the same content of the main website and the mountain biking website. One of us took the task of controlling the computer that was plugged into the television. In advance of the evaluation, we had studied the 21 heuristics from Petrie and Power (2012), so that we could prepare for the evaluation. The purpose of the Heuristic Evaluation was to examine the website and to understand the advantages and disadvantages, so that we could prepare for the Card Sorting and Think-Aloud. The Heuristic Evaluation was to be used as a preliminary study into the two websites.

Data Collection Method

To collect the data from the Heuristic Evaluation, each one of us had our computers in front of us, and were logged into Google Drive[19], where a collaborative document had been prepared with the 21 heuristics from Petrie and Power's (2012) Heuristic Evaluation Guideline. The heuristics, and the

[19] https://drive.google.com/drive/

following explanation for each heuristic, had been written into the document preliminary to the evaluation. We then entered notes into the document for further use in the Card Sorting and Think-Aloud, and could collaborate and discuss the 21 heuristics, while we at the same time could enter our notes for each of the 21 heuristics. The results from the Heuristic Analysis can be found in appendix 08.

Deriving Test Data Results

The data we collected from the Heuristic Evaluation was the Google Drive document containing all the notes from the experts. The document was then used as a starting point for the label making for the Card Sorting. The tenth heuristic "Provide clear labels and instructions" as an example were of excellent use, when we made the labels for the Card Sorting, as we could go back and see, what the notes were for the labels in the Heuristic Evaluation. The preliminary work for developing the tasks for the Think-Aloud was derived from all the heuristics, as they combined gave an overview of the advantages and the disadvantages of the two websites.

4.4 Card Sorting

Process and Setting

The Card Sorting tests were all conducted in a secluded environment without any other people present – either in e-Learning Lab's Design Lab at Aalborg University[20], an unused meeting room at Feriecenter Slettestrand or our own/the participant's apartment – depending on what was the best choice for each test. Since we tested two websites, we created two Card Sorting tests that were similar in design and participant tasks, but used two sets of cards (A-cards and B-cards). The A-cards were used for testing the non-mountain bike website where the participant requirements are less defined than in the mountain-bike domain-heavy website, for which we used B-cards. The A-cards consists of a total of 43 cards, and the B-cards consists of 31 cards. The cards can be seen in appendix 02.

Each test part (A and B) consisted of four identical tasks (see appendix 01 for more details), making the entire progress an eight-step test for each participant:

1. A sorting with no restrictions (unlimited number of groups, cards in each group etc.).
2. Condense the groups from task 1 into a maximum of five (and minimum three).
3. Use a card from each group to be the group's overall label. The participant was also allowed to create his own label for this, if he wished.

[20] http://www.ell.aau.dk/

4. Are any cards (from any group) something the participant would like to see on the front page of a website (after being told the context) and are any cards deemed redundant or unnecessary?

These tasks were each created to serve a specific purpose; the first tasks investigate the mental model of the participant by not providing any context for the test, and no restrictions for how the cards were sorted. The second and third task investigate how the participants would sort the cards into three to five categories, and how they would label that category, exemplifying a subjective, but possible way to sort the cards' labels on a website's menu or navigation system. The last task helps to see if some cards' labels are more important to the participant, and if there are labels that are out of context or useless.

After having completed 20 tests, we had collected enough empirical data to be able to draw conclusions based on the statistics of the card sortings. Combining those statistics with Information Architecture theory also helps us to determine which labels were most useful, both from the users' perspective, but also from a perspective based in the Information Architecture theoretic knowledge.

Data Collection Method

To collect data from the Card Sorting tests, we documented every step in the process by taking photographs of each step that every participant went through (primarily photographs of the participants' sorted cards), ending up with a total of over 160 photographs (20 participants x 8 tasks) of every participant's cards and how they were sorted. For each photograph, we included a paper-slip where we could cross of which test this was a photograph of, making sure that would not mix up the photographs later in the analysis.

The paper slip also included the number of the participant, so that we would easily be able to identify which participant did each card sorting, as it is important to differentiate between the card sortings done by domain experts and domain novices. The participants are all kept anonymous, but we made sure to keep track of them by name, age, domain expertise and by assigning them a random number before each test in a private document.

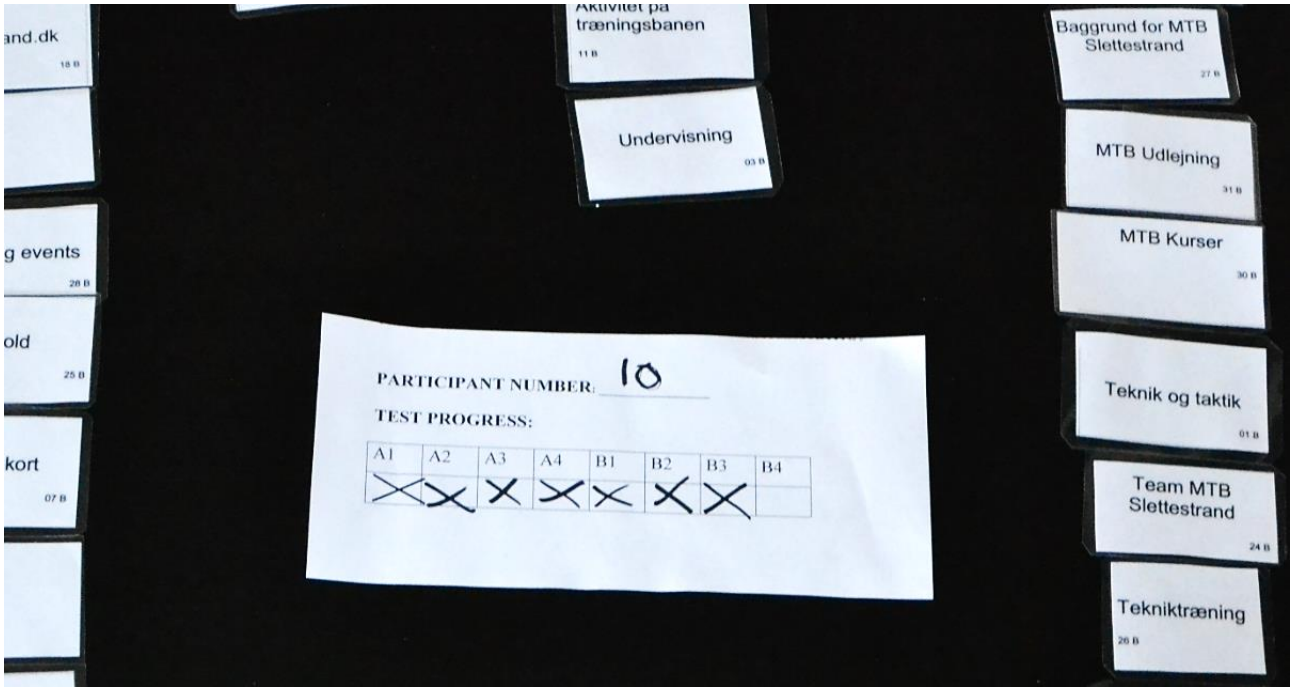

Figure 2 - Every card sorting would include a slip including the participant's number and how far in the test progress the sorting was from. It is also possible to see how each card has its own number and if they are cards from the A or B test for easy analysis. The photograph above is from task B3 (the 7th task/B cards) for participant number 10. On the cards, the unique numbers used for analysis can also be seen under each card's label.

We were also with the participants during the entire process, taking notes on our laptops if there was something that we could not capture on a photograph, or if a participant had any questions or interesting and useful comments during the tests. We did not interrupt the tests, and to make sure that the participants would have equal opportunities in terms of time, we created a time estimation for each task, limiting the entire Card Sorting process to about 30-35 minutes for each participant. This is time enough to complete all the sortings, and also ensured that the participants would not take too long, overthinking how they sorted their cards. There was never any need to stop a participant in the process for taking too long, however.

Deriving Test Data Results

As mentioned, the Card Sorting tests provided us with more than 160 card sortings that needed to all be considered in a larger perspective. In the theory section concerning Card Sorting, we already described how we were going to use Donna Spencer's Excel spreadsheet for analysing Card Sorting results (Spencer, 2009, p. 181). After having completed all the sortings, we input all the data from each sorting into one of four copies of that spreadsheet, as we differentiated the sortings into four categories (see appendix 04 for the filled in spreadsheets containing all the raw sorting data):

- Domain Expert A-cards
- Domain Expert B-cards
- Domain Novice A-cards
- Domain Novice B-cards

The spreadsheets contain every card sorting that we completed, and they show some of the statistical tendencies that is present in the four different sortings. They also provide information on which labels were most often chosen to represent each category of cards, the number of unique cards in each category and how many participants used the same labels for each category. From these spreadsheets, we also extracted data on which cards were chosen by the participants for category labels, which cards they deemed redundant or unnecessary, and which cards they found important enough to feature on a website front page (see appendix 07).

4.5 Think-Aloud

Process and Setting

As explained in the section concerning the Think-Aloud method, asking a person to "think aloud" can feel very unnatural to that person, but it is necessary and an essential part of the method. To ensure that the participants were prepared for thinking aloud, for each participant we demonstrated how we would think aloud (by giving a two-minute demonstration of one of us thinking aloud while using a random website), and then by giving each participant a task where they had to put staples into a stapler, commenting or thinking aloud on the process while they were solving the task. Doing so, we could then proceed with the test, or give the participant suggestions for how to think out loud more confidently or efficiently (Barnum, 2011, pp. 205-206).

The setting for the Think-Aloud tests were the same as with the Card Sorting tests; conducted in a controlled setting, either e-Learning Lab's Design Lab at Aalborg University, an unused meeting room at Feriecenter Slettestrand or our own/the participant's apartment.

After the participant felt ready to begin the test, we provided a context for the participant to take into consideration when completing the tasks. One of us sat beside the participant to observe the participant and his actions, prevent stalling by being ready to help or ask follow-up questions if

necessary[21], while the other would be taking notes from farther away, trying not to distract the participant and the test.

The participant was then asked to complete a set of tasks on the two websites. The tasks were composed by taking into account some of the issues that were found during the heuristic evaluation, but also consisted of regular tasks that would be completely normal for a regular visitor of the site to complete, such as finding the price for various services, or finding information on a specific mountain biking track. If the participant at any point gave up on a task, or made it clear that he would never have spent so much time on a task in a real situation, we would continue to the next task instead – the participants were instructed to do so beforehand.

We had also made sure that the tasks investigated, or could be used to investigate, the different attributes of usability, which we defined earlier in the usability theory section:

- The usefulness, effectiveness and efficiency
- Learnability and satisfaction
- Errors and safety
- Intuitive design

Some of the tasks were designed as open questions, inviting the participant to explore and use the websites as organically as possible, other were interview-style questions where the purpose was to inquire into the participant's subjective conception of the sites and his satisfaction of them. After each Think-Aloud session, we asked a few follow-up questions on how the participant would change two things on each site, and if the content they had seen lived up their expectations of the content on websites like those we were testing. The context for the tasks, and the tasks themselves can be found in appendix 03.

Data Collection Method

For collecting data during the Think-Aloud tests, we designated one of our laptops as the device used for the participants to use while thinking aloud. The laptop was configured with a mouse (so that the participant could choose between using a mouse or the laptop's built-in trackpad dependent on

[21] As described in the section concerning the method behind the Think-Aloud process, it is primarily up to the participant to do "the talking", but not all participants are equally comfortable completing this process, and it would sometimes be necessary to ask interview-style questions to keep them thinking and verbalizing their thoughts and prevent stalling.

personal preference and level of comfort) and an external microphone to record the participant's verbalised thoughts and the interview questions + tasks.

The laptop was configured with QuickTime software, which was used to record what happened on the computer's screen (the software recorded what the participant saw, indicated when and where he clicked, his mouse movement, keyboard inputs etc.). It also recorded voice and audio from the external microphone, and compiled it down to a video which could be used for analysis later.

Deriving Test Data Results

The data we collected during the Think-Aloud tests consist of 10 video files of varying length (shortest: ~15 minutes, longest: ~36 minutes) depending on how fast the participant completed all the tasks. The data also includes our field notes for each participant. After having completed all the tests, the video files were watched, and additional notes were taken. The second notes were important, because as the videos were watched again, it was much clearer to see what the participants actually did (where did he click, what did he focus on etc.), and if these actions were in contradiction or agreement with what was being verbalised. These notes also include the time taken for each task (in minutes and seconds), and the number of mouse clicks used to complete each task. We also noted if any tasks were skipped or given up on. These notes can be seen in appendix 06 and the 10 Think-Aloud videos can be found in appendix 09.

4.6 Summary and Reflections on the Testing Process

One of the things that surprised us the most was how much time it takes to prepare and schedule all the tests with the test participants, and how little of that time that was actually spent on testing. We estimated, that if we had been able to complete all 20 tests in rapid succession, we could have completed the entire process in two or three days. It ended up taking around three weeks.

What surprised us also, was the testing process itself. In most cases, we were very impressed with how much energy the test participants put into the tests, and how engrossed they were in the tests when they first got started. From a usability evaluation standpoint, it was very impressive to observe patterns emerge when observing the participants solve the same set of tasks, and the variety of ways the participants managed to solve (or not solve) the tasks. Many of the participants, after being introduced to the purpose of the tests, were also very positive on the idea of evaluating usability in order for a possible redesign or rework of some of the websites' elements, and were all able to

verbalise or signal at elements, features or functions that they thought needed improvement in another design iteration.

Some of the issues we encountered during the process have also influenced how we would change the test process if we had the possibility of doing it over again:

Firstly, we had prepared the tasks for each test, but especially in the Think-Aloud tests, it was sometimes hard to continue to the next task with fluid transition, as that type of test is very subjective. Sometimes, the participants had already solved one or two of the next tasks unconsciously, making the flow of tasks more irregular and changing from participant to participant. This issue does not influence the test data much, it just makes job of measuring the time and mouse clicks spent on each task more difficult later in the analysis.

Secondly, if we had known how time-consuming the testing process would be, we would probably have planned the entire process much earlier. This includes the participant recruitment process, which we were finishing after we had begun testing with the first participants. In the way that we did it, there was not much room for mishaps or big changes to the plan, because of our limited time schedule. Luckily, we only had one person not showing up for the tests, and we had a few backup participants ready to help if necessary. This is fortunate, especially when considering no-show rates for usability studies seems to lie approximately around 10-11% (Barnum, 2011, p. 161; Sova & Nielsen, 2003, p. 7) and we had a total of 20 participants with a no-show rate of 5%.

These aspects are all important parts of the usability tests, and the patterns that emerged in the two participant groups, both in the Card Sorting and Think-Aloud results, as they indicate where on the two websites there are serious usability issues, but also where the participants found the Information Architecture elements to be in compliance with their mental model and behaviour models, which can illuminate some of the positive usability aspects.

It is also worth noticing, that all of the data we will present next, will be used qualitatively. The research methods used to collect them, especially the Card Sorting and Think-Aloud methods, yields data that can be used both quantitatively and qualitatively. We have decided to approach the data from a qualitative standpoint, as we will focus on the individual participants' data, rather than trying to conclude ideas from a quantitative standpoint.

4.7 Data Presentation

In this section, we will present the data that we have derived from the respective tests that we have completed that also involves test participants; the Think-Aloud tests and the Card Sortings. The data that we chose to present here will consist of the most essential test results from the usability evaluation context, and will be the data that we will primarily use to answer our hypothesis concerning the influence of test participant domain expertise in usability evaluations, and will also be useful for a usability evaluation of Feriecenter Slettestrand's websites. This means that we will omit some of the data, as there would otherwise be much redundancy or irrelevant data, simply because the quantity of the data we have collected during the before mentioned tests is very large (160+ photographs of card sortings, approximately five hours of Think-Aloud session videos, Heuristic Evaluation results and notes for all three tests). This also means that some of this data will be presented in visual form, rather than raw data results, making the data easier to understand and read. However, all of the raw data and notes can be found in the appendixes.

4.8 Data from the Card Sorting

The list of all cards from the card sortings (labels used and the individual card numbers) can be found in appendix 02 for reference, but in most cases, when referring to a specific card, we will write for example "B05 Nyheder", which shows if the card is from the A or B part, which number it is and what the label for that card is.

Much of the data from the Card Sorting will be presented in graphs where the X axis consists of the cards and the Y axis is the number of times that specific card was used by participants in the Card Sorting task concerned. Since there were no requirements for how many categories the participants had to create for each task, the number of cards for each task and user group (domain novices and experts) varies. The graphs are sorted by A and B cards, user group for each card type, and a total (results from domain novices and experts in total) for each card type. The data that was used for these graphs can be found in appendix 07.

Task: Cards for the Front Page

Figures 3 - 8 depict the results from the task where the participant could choose any cards, when asked if there were any cards that the participant deemed so important that it deserved a spot on the front page of a website. Figure 3 shows how the domain novice participants chose A-card labels that were deemed so necessary, that it needed a position on the front page of the website. These labels indicate which type of information the participants felt was most important or relevant in the context of Feriecenter Slettestrand's main website. The "Other" column consist of the cards that were chosen only once by the novice participants.

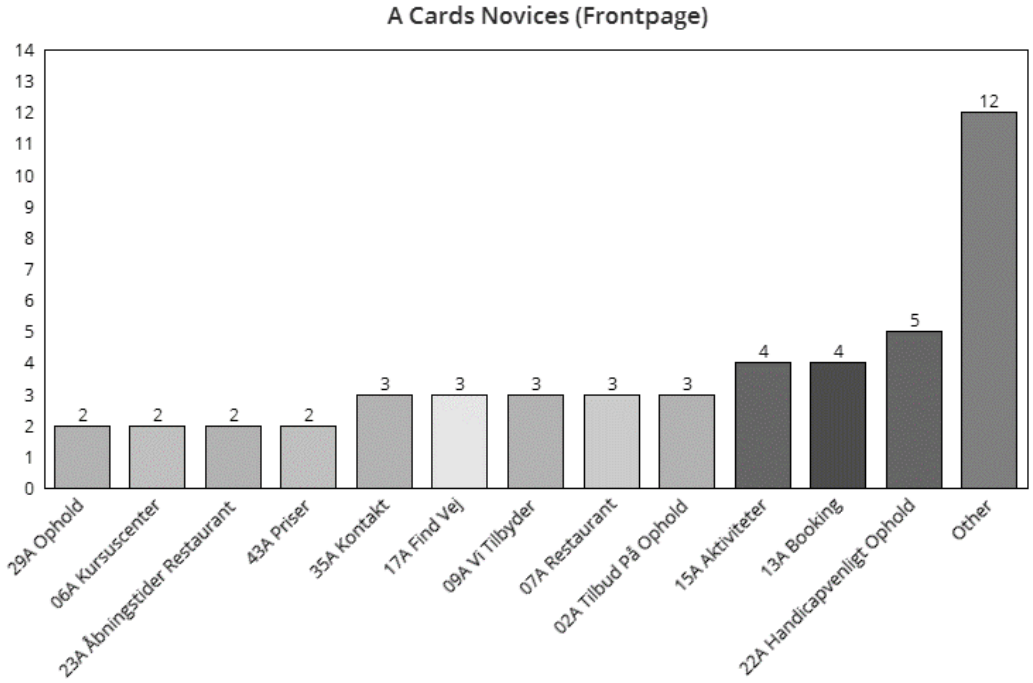

Figure 3 – A Cards Novices (Frontpage)

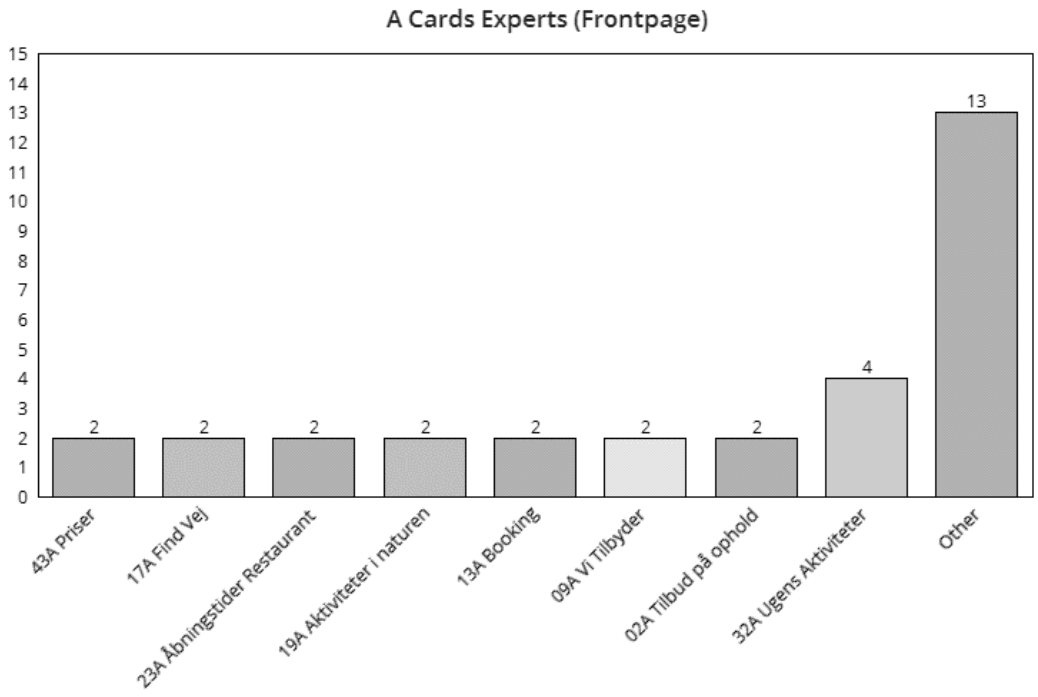

Figure 4 – A Cards Experts (Frontpage)

Figure 4 shows how the expert participants have chosen A-cards for the front page of the main website. In comparison to figure 3, the experts agree less than the novices, as they have only one card with more than two participants agreeing ("32A Ugens Aktiviteter").

Figure 5 shows the dispersion of A-card labels that were selected for the front page of the main website. The data from figure 3 and 4 are added together, showing the results of the task independently on user groups. The most agreed on labels are "13A Booking" and "22A Handicapvenligt ophold", which each had six of the 20 participants choosing it. It is also worth noting that there are 14 labels ("Other" column) that were only chosen once.

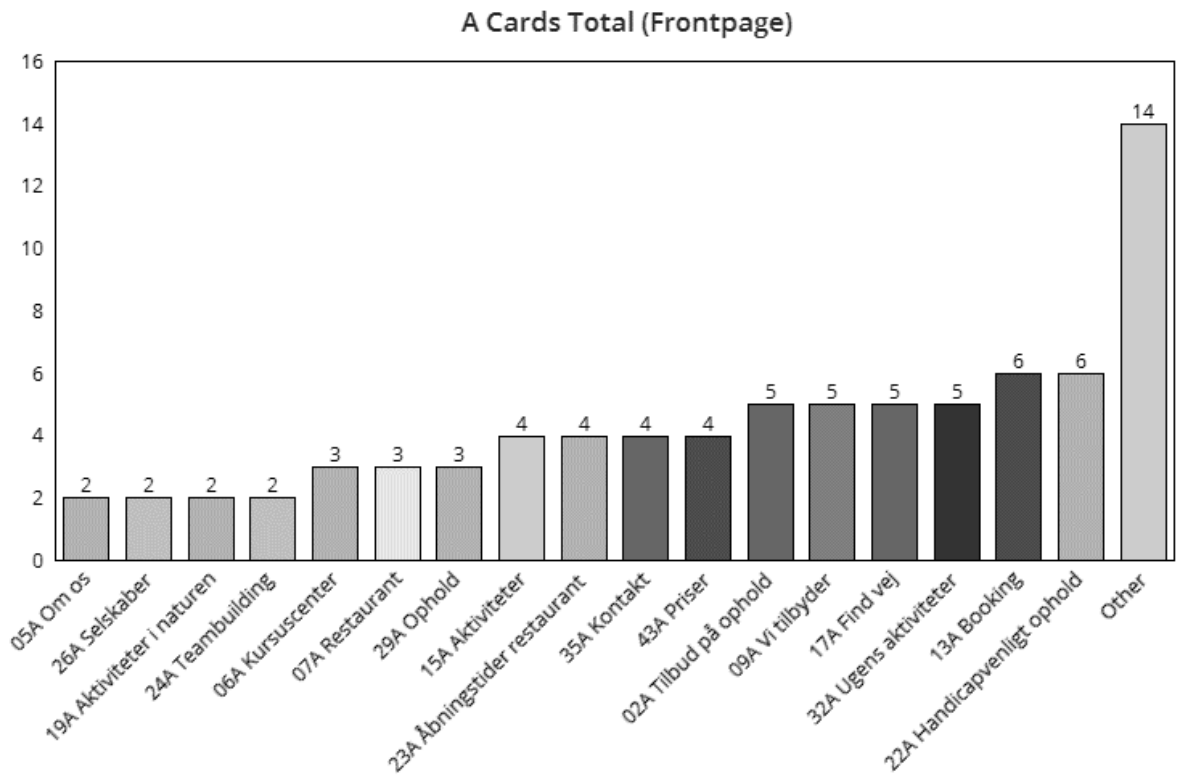

Figure 5 – A Cards Total (Frontpage)

Figures 6 - 8 depict the results from the same task as figure 3 - 5, but completed with the B-cards, which were focusing on the mountain biking domain. In figure 6, the novices' results are shown, and figure 7 shows the experts'. Compared to how the experts did not agree much in figure 4, they agree more on the labels in the B-cards, than they did with the A-cards. The novices have a higher count of total labels too, indicating more disagreement on between the novices, than the experts.

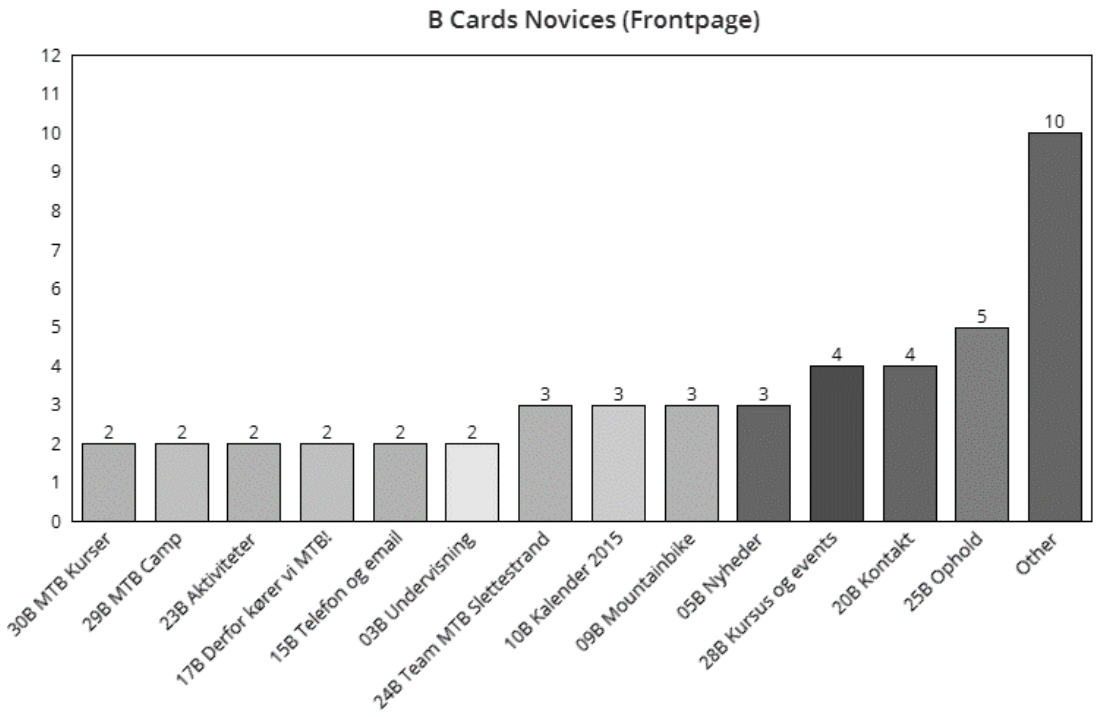

Figure 6 - B Cards Novices (Frontpage)

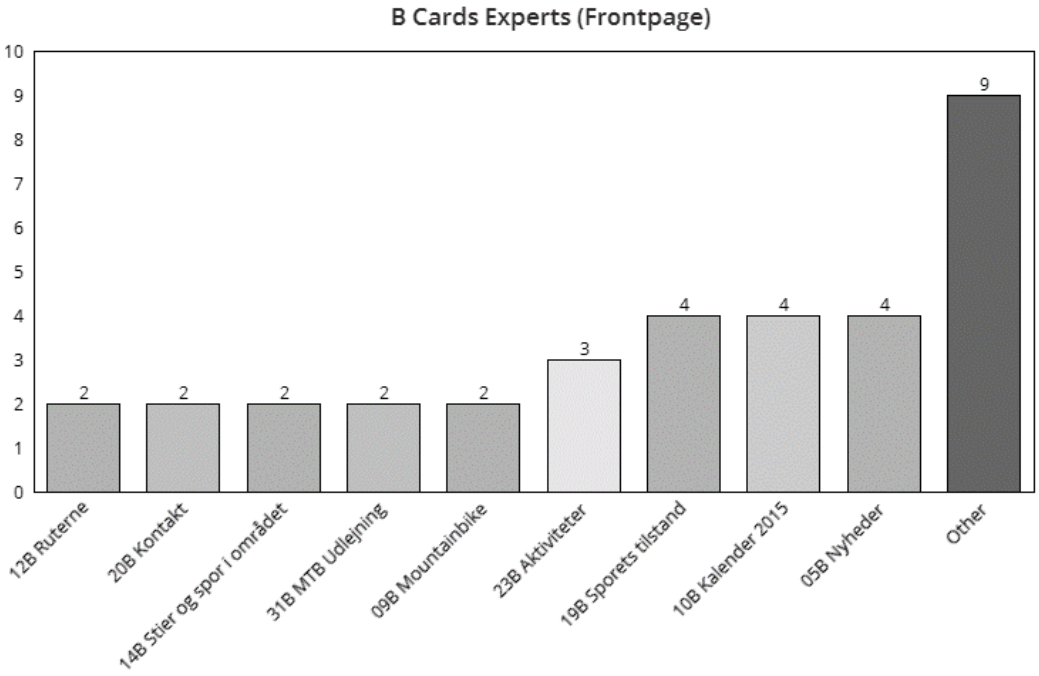

Figure 7 - B Cards Experts (Frontpage)

Figure 8 shows the total results from combining the results of figure 6 and 7, indicating that the B-card labels that were most agreed on for the front page, between all of the participants, are “20B Kontakt”, “25B Ophold”, “05B Nyheder” and “10B Kalender”.

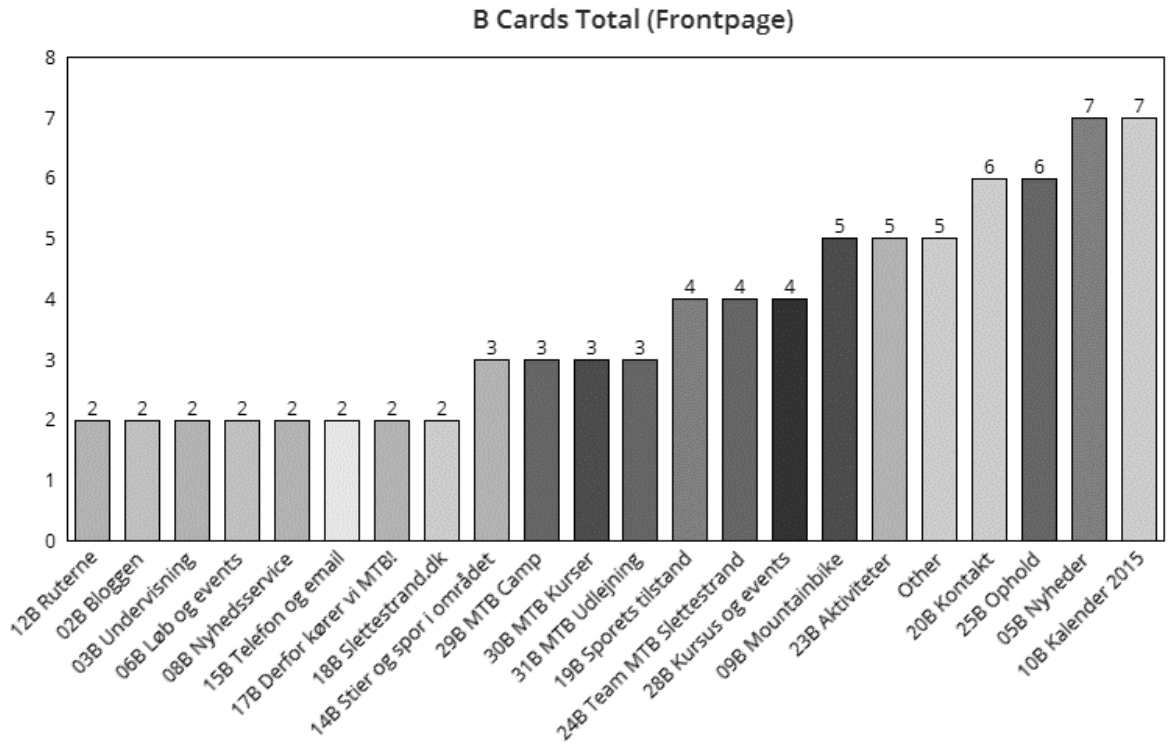

Figure 8 – B-Cards Total (Frontpage)

Task: Redundant or Unnecessary Cards

Figures 9 - 14 depict the results from the task where the participants were asked if there were any cards that they did not deem necessary for a website like Feriecenter Slettestrand’s. It is worth noting, that a lot of A-cards were deemed unnecessary by a single expert participant, which can explain the large number of single cards depicted in the “Other” column of figure 10. Besides from that, neither the novices nor experts chose many A-cards that they thought of as unnecessary or redundant. The experts chose a few more than the novices though.

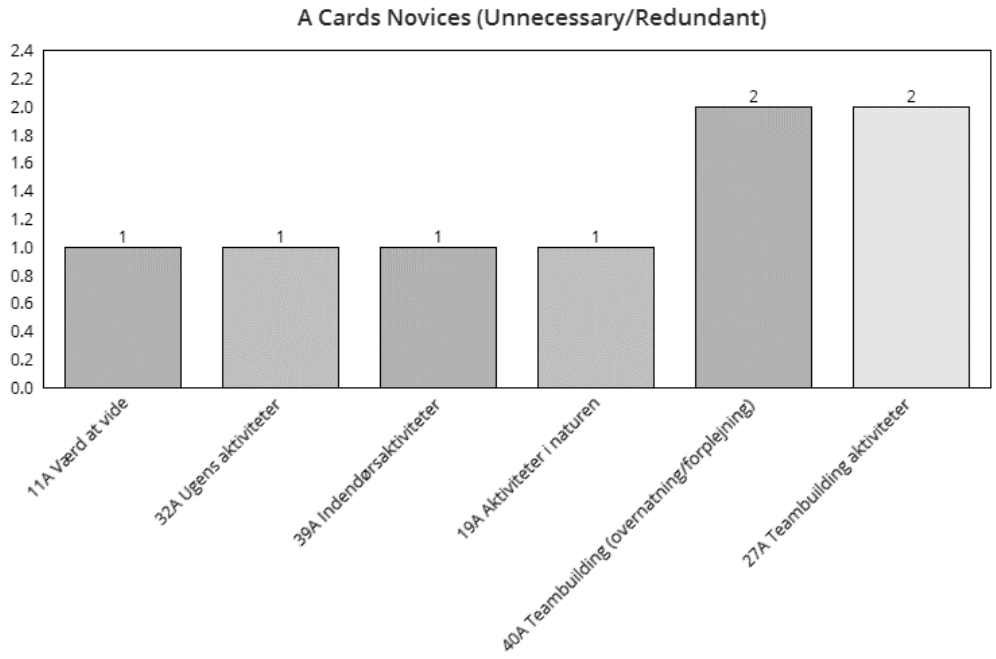

Figure 9 – A Cards Novices (Unnecessary/Redundant)

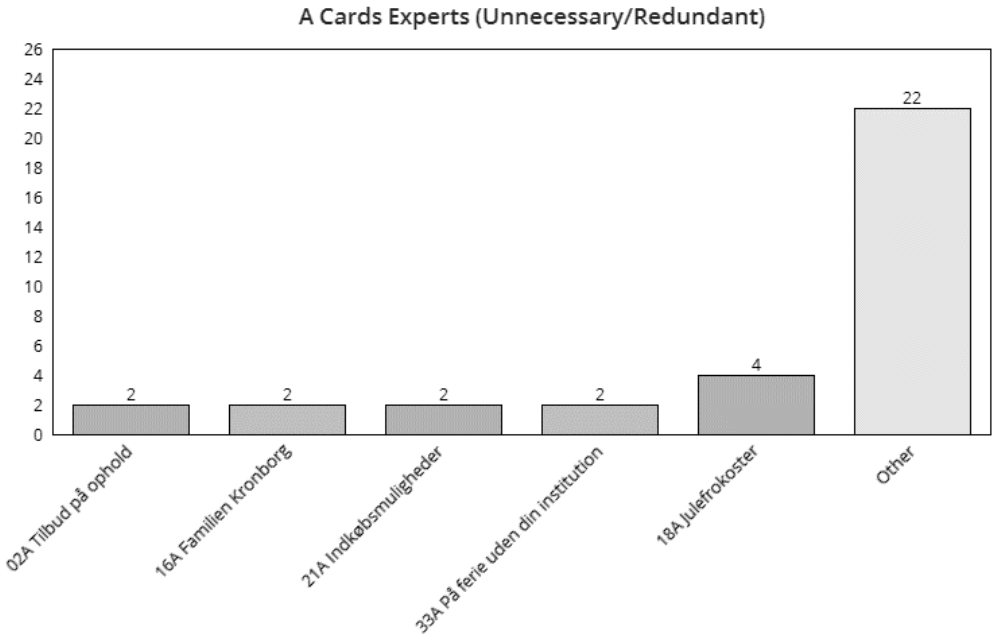

Figure 10 – A Cards Experts (Unnecessary/Redundant)

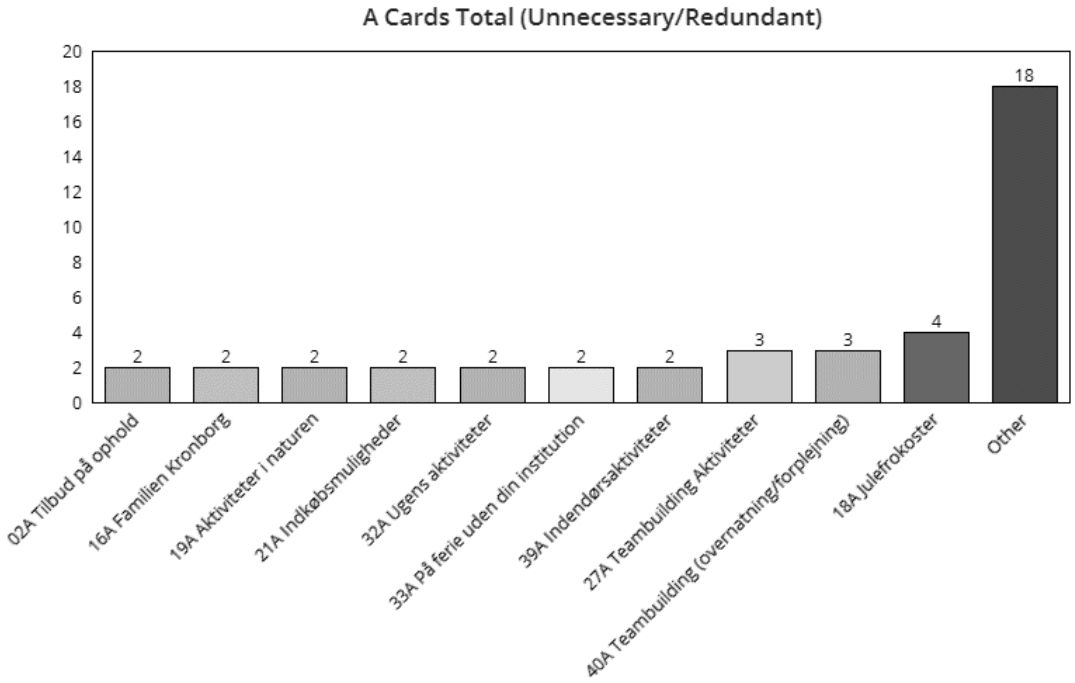

Figure 11 – A Cards Total (Unnecessary/Redundant)

In figure 11, the total number of A-cards chosen as unnecessary or redundant between the two user groups is summed up. There are not many surprises, except for the large amount of cards that the one expert participant chose. "18A Julefrokoster" is the most chosen card, probably because of its relevancy to a specific season, which made it stand out. The next two most agreed on are both concerning teambuilding activities ("40A Teambuilding (overnatning/forplejning)" and "27A Teambuilding aktiviteter").

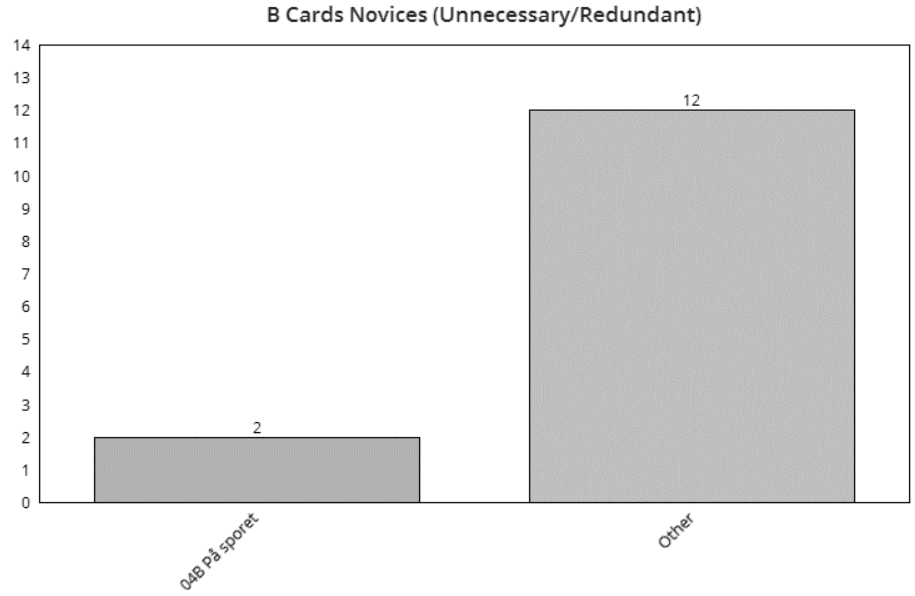

Figure 12 – B Cards Novices (Unnecessary/Redundant)

Figure 12 is quite interesting, as it shows how the novice participants chose B-cards that they felt were unnecessary or redundant. “04B På Sporet” is the only card that was chosen more than once, but there were 12 other B-cards chosen only a single time, indicating that the novices disagreed a lot on which B-cards they felt were unnecessary or redundant.

Figure 13 shows how the experts chose B-cards for the same task, which they agree much more on, than the novices. The experts chose more cards, but more participants agreed on the same cards.

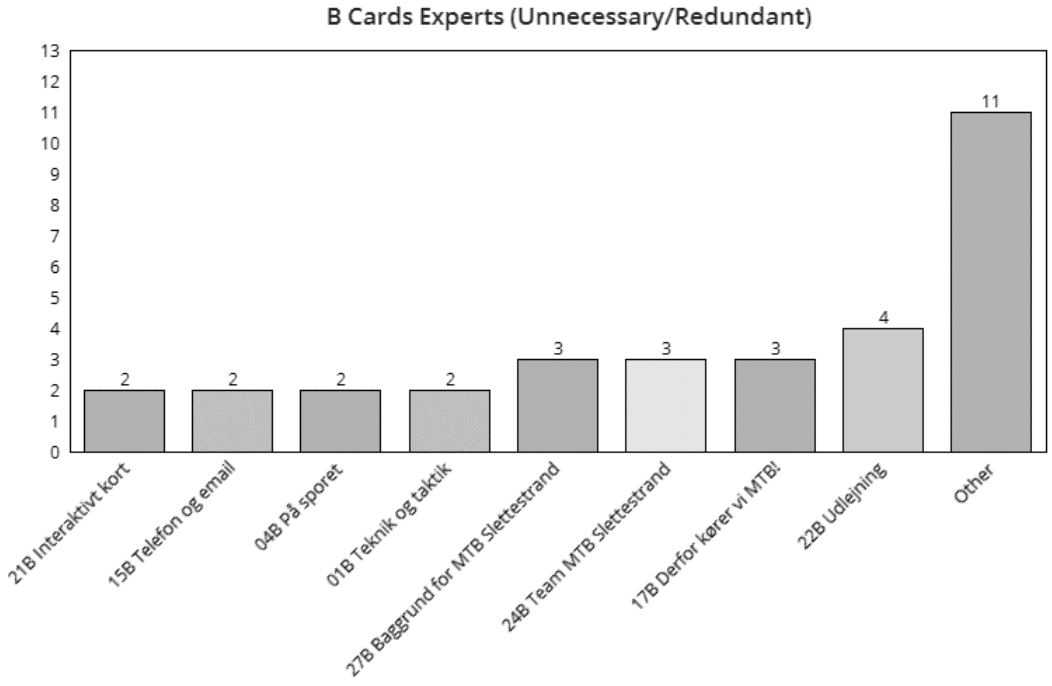

Figure 13 – B Cards Experts (Unnecessary/Redundant)

Figure 14 shows more or less the same as figure 13, mainly because the novices were so disagreeable that the data from figure 12 does not influence the data in figure 14 much. There are a few cards that more than three participants agree are unnecessary or redundant, most notable one being “22B Udlejning”. The possible reason for this card being the most agreed on (with five participants total) is that that it is very similar to another card (“31B MTB Udlejning”), and thus might feel redundant.

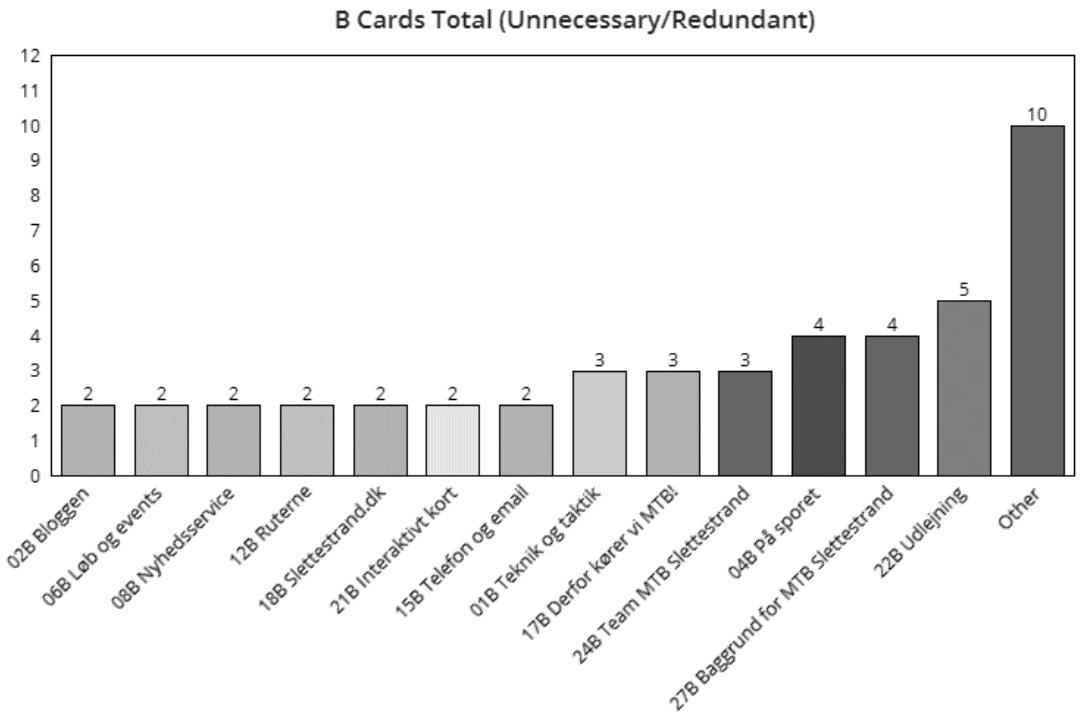

Figure 14 – B Cards Total (Unnecessary/Redundant)

Task: Category Labels

Figures 15 – 20 show the results from the task where the participants were asked to choose one card from each of the categories that they had sorted out, and make that card the label or "headline" for that category. As described in the research design section, the participants were also allowed to create their own labels for this task. However, no two homemade labels were made twice, and thus are the homemade labels counted into the "Other/Participant created label" columns.

In figure 15, it is clear how the novice participants are very agreeable on which A-cards to use as category label; One card was chosen as a category label by 5 out of 10 participants ("04A Om os"), while two other cards were chosen by 7 out of 10 ("15A Aktiviteter" and "07A Restaurant").

The expert participants have chosen less cards, and are also less agreeable overall, as seen in figure 16, but they still have two cards that were chosen 6 out of 10 times ("09A Vi tilbyder" and "04A Om os"). All of the expert participants (10 out of 10) have used card "07A Restaurant" as a category label, which is also one of the cards that the novices used the most times (7 out of 10). Card "04A Om os" and "15A Aktiviteter" are also popular in both participant groups.

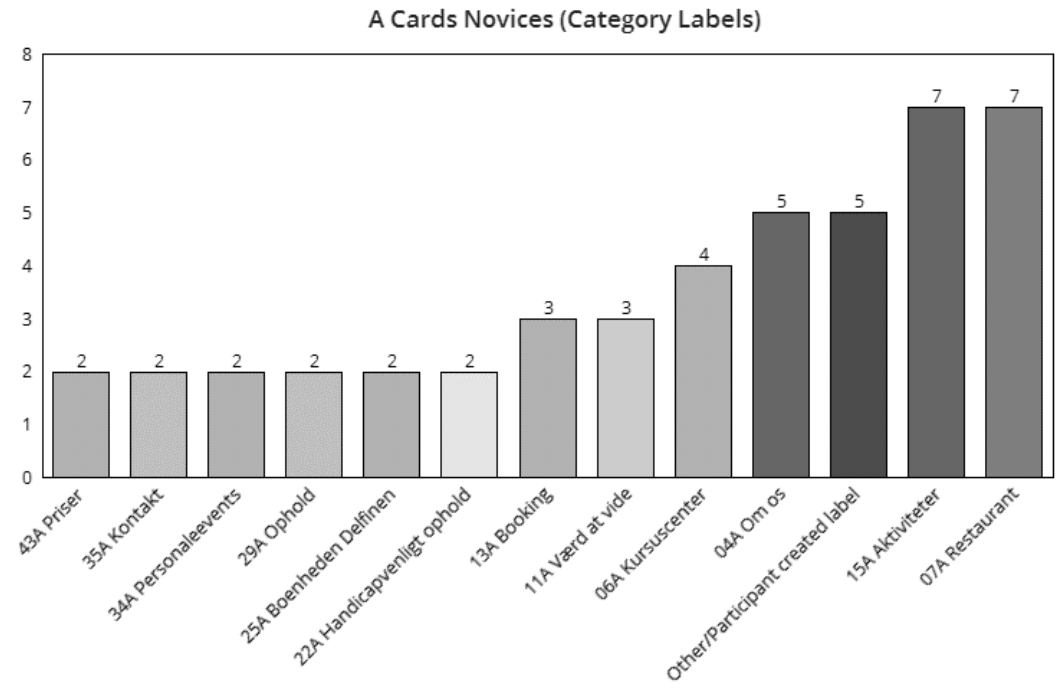

Figure 15 – A Cards Novices (Category Labels)

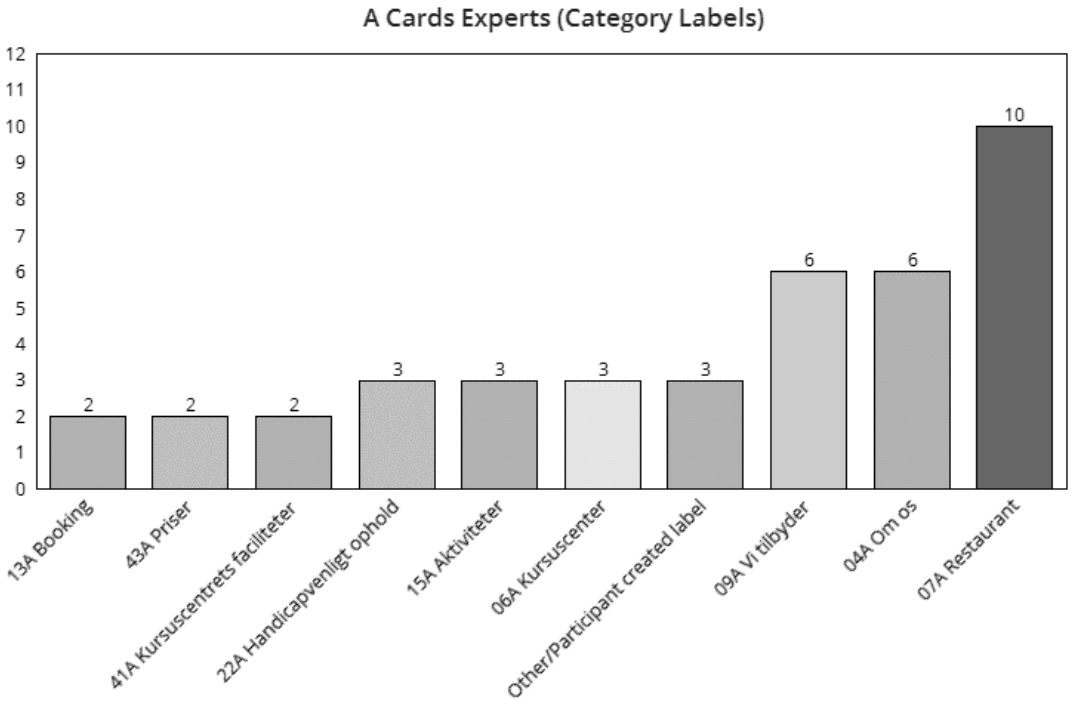

Figure 16 – A Cards Experts (Category Labels)

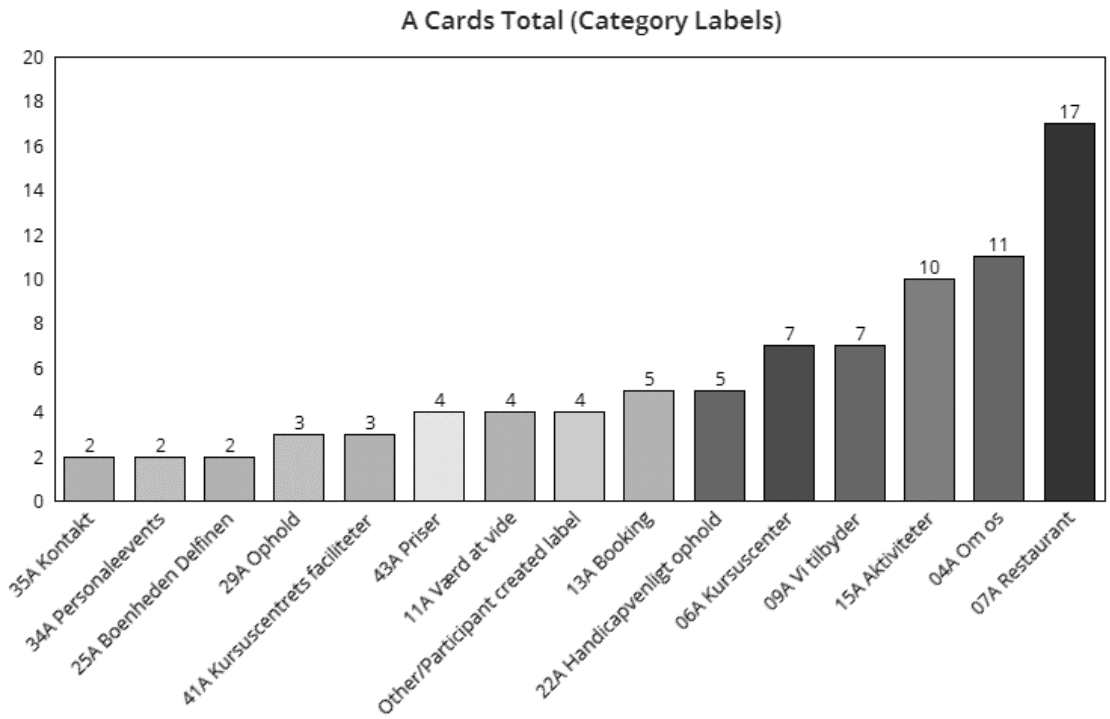

Figure 17 – A Cards Total (Category Labels)

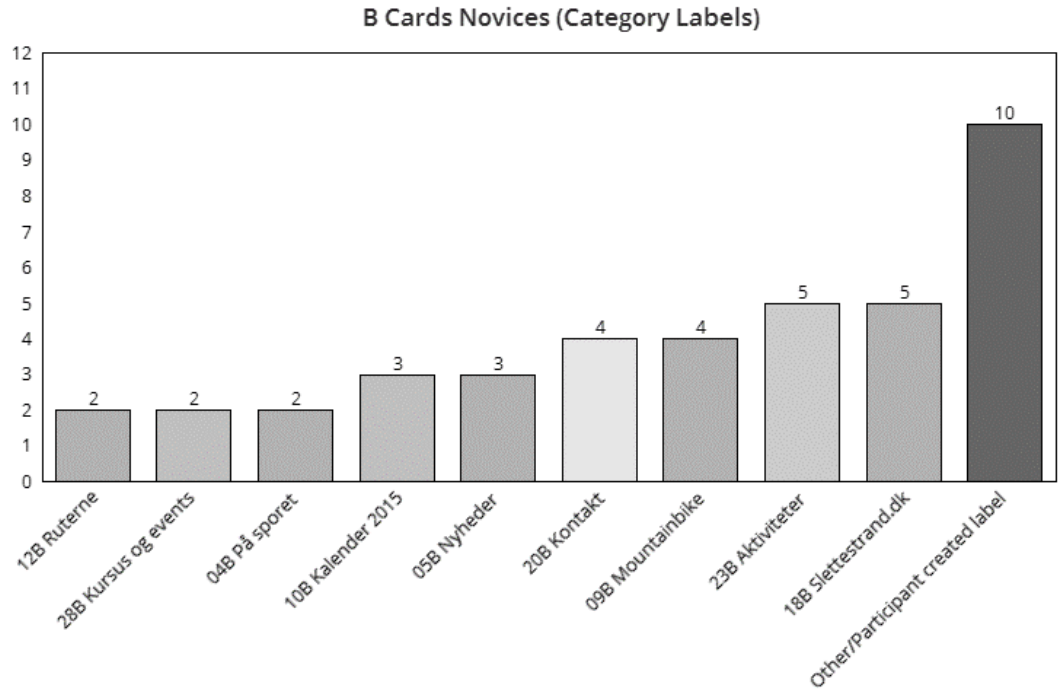

Figure 18 – B Cards Novices (Category Labels)

Figure 17 shows clearly what cards were most often used as category labels by all of the participants, but there are not many surprises besides what we already saw in figure 15 and 16.

Figure 18 shows the novice participants' results of the same task, but with the B-cards. There is a good level of agreement between the top four cards "20B Kontakt", "09B Mountabike", "23B Aktiviteter" and "18B Slettestrand", but there is also quite a high number of cards that were only chosen three times or less, especially when taking the "Other/participant created label" column into account.

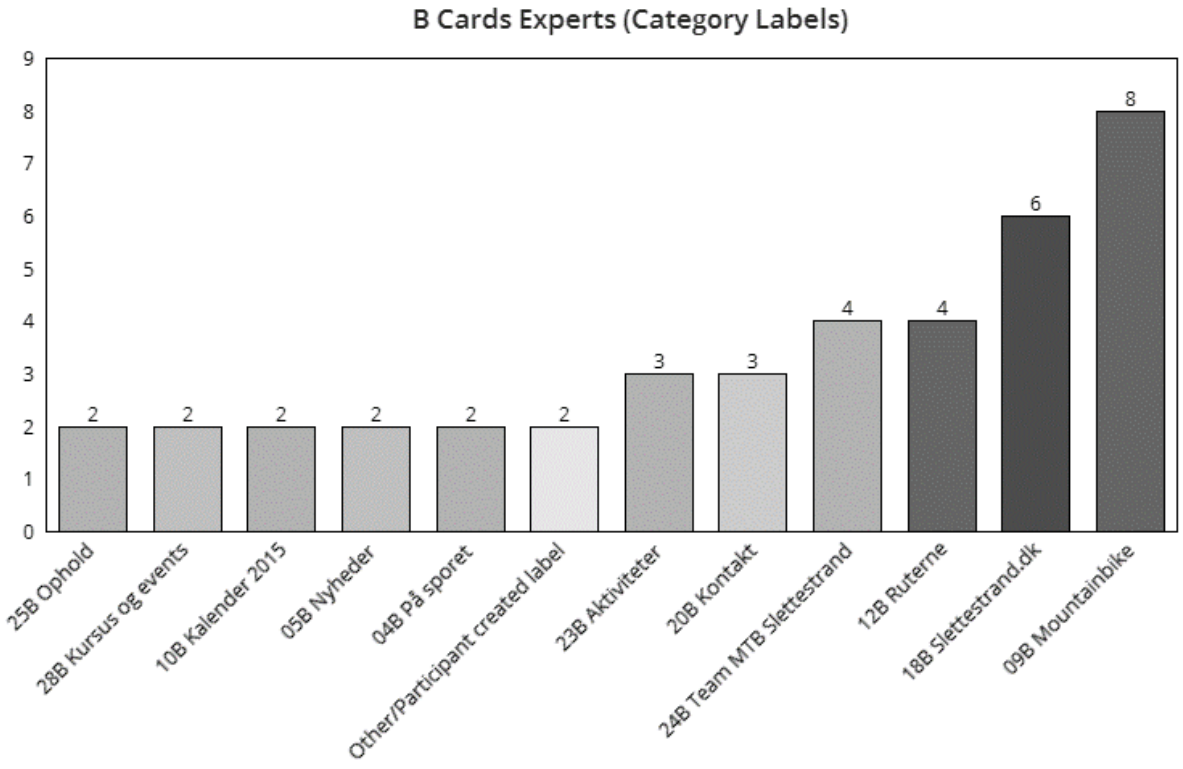

Figure 19 – B Cards Experts (Category Labels)

In figure 19 it can be seen that 8 out of 10 expert participants has chosen card "09B Mountainbike" as a category label, alongside with other mountain biking relevant cards as the most popular ("12B Ruterne" and "24B Team MTB Slettestrand") indicating that they have put more focus on the different topics within the mountain biking domain than the novices did in figure 18.

In the last figure, number 20, we can see that the most popular card for labelling card categories is "12B Mountainbike" closely followed by "18B Slettestrand.dk". Those two cards are very general, and it is understandable how participants have used those two often. We can also see that activities and calendars/news has been chosen for category labels quite often.

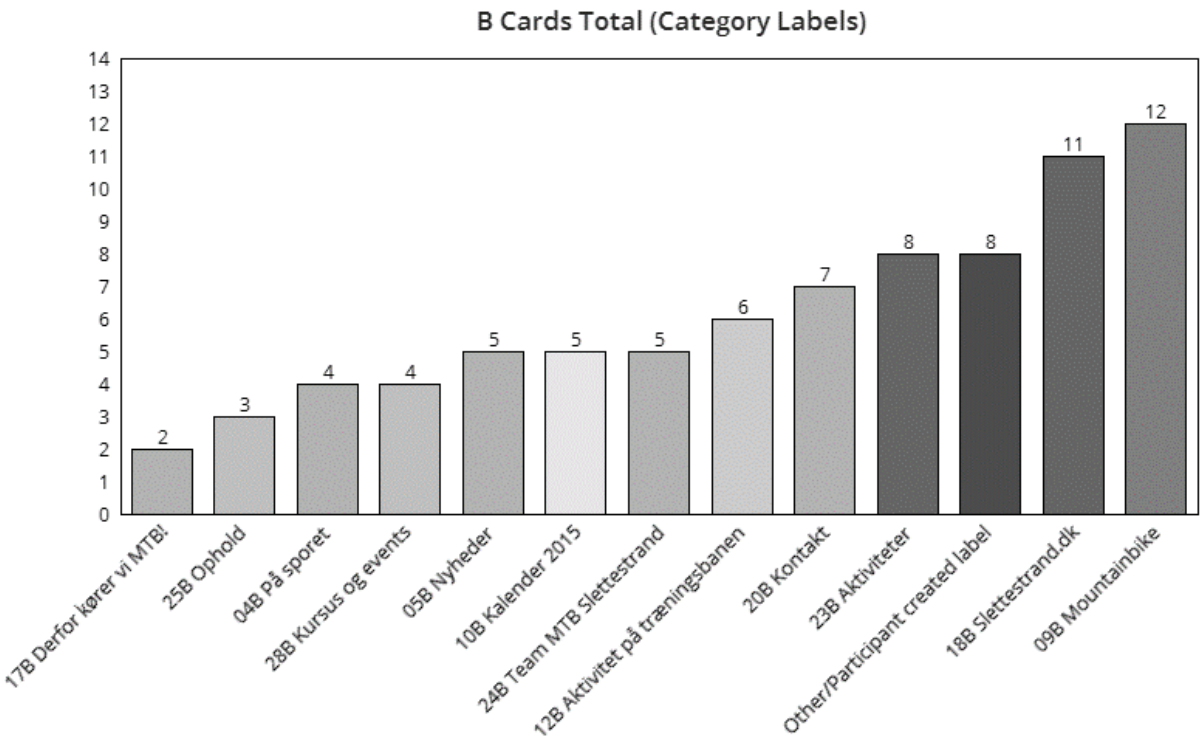

Figure 20 – B Cards Total (Category Labels)

4.9 Raw Card Sorting Data and Statistics

As mentioned in the last section, all the data from all the Card Sortings that we completed is filled into four spreadsheets (appendix 04). Each spreadsheet contains all the categories and labels that were used (including the ones that participants created – see the *CatsRaw* tab in each spreadsheet), the number of unique and total cards for each category, and how many participants who used each category (*CatsSummary* tab), and most interesting; a visualization of the statistics of all the cards in percentage (*Correlation* tab). Each spreadsheet also contains all the raw data from each card sorting (tabs *Sort1*, *Sort2*, *Sort3* etc.). Since there is too much data to visualize here, we will refer to these spreadsheets and their tabs later in the analysis, rather than copying all the data into here. We have chosen to do so, as we will not be using *all* of the data from each spreadsheet in the analysis anyways, because there is data with low relevancy to the problems that we are investigating.

4.10 Data from the Think-Aloud

Driving out data from the Think-Aloud sessions is more nuanced, than with the Card Sortings. Rather than being almost pure statistics (like the Card Sorting data), this data is obtained by looking at the test results from two perspectives; a statistical perspective that takes into account the raw data from the tests (time spent on each task, number of mouse clicks needed, how many tasks were failed etc.), and a more general perspective based on participant actions and verbalized thoughts combined with our notes, findings and observations by watching the sessions again on video. This process is inspired by how Tullis and Albert (2013) recommend to combine and compare usability metrics (Tullis & Albert, 2013, pp. 187-208).

In the process of naming the participants, the participants were assigned a random number. The given number was to make sure that the participants were anonymous. The order in which the participants participated was random and that resulted in a non-linear naming of the participants participating in both Card Sorting and Think-Aloud. The novices participating in both Card Soring and Think-Aloud are therefore named Participant 1, Participant 8, Participant 10, Participant 12 and Participant 18. The experts are named Participant 2, Participant 4, Participant 7, Participant 11 and Participant 14. This does however not give a good overview, when referring to them in the data and later in the analysis. The novices Participant 1, 8, 10, 12 and 18 are therefore referred to with an [N] for “Novice” and a number from 1-5, as there are five novice participants, as shown below:

- Participant 1: [N1]
- Participant 8: [N2]
- Participant 10: [N3]
- Participant 12: [N4]
- Participant 18: [N5]

The same is done with the expert participants Participant 2, 4, 7, 11 and 14, who are now referred to with an [E] for "Expert" and a number from 1-5 as there are five participants, which is shown below:

- Participant 2: [E1]
- Participant 4: [E2]
- Participant 7: [E3]
- Participant 11: [E4]
- Participant 14: [E4]

4.11 Statistical Think-Aloud Data

In the following tables, we have summed up the raw, statistical data derived from the Think-Aloud session videos. The tables show how much time (in seconds) each participant spent on each of the tasks (appendix 03) they were asked to complete. It also shows how many actions (mouse clicks) were used for each task and how many of the tasks were failed by each participant (if a task was failed, the time and mouse clicks for that task are in [brackets]). Figure 21 shows the data from the tests with novice participants, and figure 22 shows the data from the experts. Task 6, 12 and 13 are not included in this table, as they were interview-style tasks that did not involve any participant action besides just talking.

For each table, we have calculated the average fail percentage, time spent and number of mouse clicks for both participant groups in each task. If a task, for any reason, was skipped or the data is unavailable (see for example participant 14, task 05), the calculation for the averages has been adjusted. Tasks that were failed are still included in the calculation, as the participants were instructed to tell us if they wanted to give up on a task, or if they spent more time than they would have in a real life situation, which is why we still consider the failed tasks valid attempts. Lastly, the color coding shows if the average result or task fail percentage of that user group is lower (green), equal to (yellow) or

higher (red) than the other user group's average results for the same tasks, making it easier to compare and visualize the difference between the results of the two participant groups.

It can quickly be assessed, that the domain experts had a slightly lower fail percentage, and were generally faster to solve the tasks they were given, and with (marginally) less mouse clicks, compared to their domain novice countergroup.

NOVICES												
		Task 01	Task 02	Task 03	Task 04	Task 05	Task 07	Task 08	Task 09	Task 10	Task 11	Combined
[N1]	Seconds:	45	[85]	70	38	15	55	[83]	13	8	[103]	
	Clicks:	2	[5]	3	2	2	2	[6]	2	0	[3]	
[N2]	Seconds:	159	[192]	142	[146]	20	78	46	26	4	[19]	
	Clicks:	3	[7]	4	[6]	1	5	3	3	0	[1]	
[N3]	Seconds:	263	[240]	72	99	83	79	48	7	35	36	
	Clicks:	5	[5]	2	4	4	3	2	1	0	1	
[N4]	Seconds:	147	26	45	31	21	59	52	17	7	24	
	Clicks:	5	1	2	2	1	2	1	2	0	2	
[N5]	Seconds:	[121]	19	41	[420]	23	87	65	[87]	6	25	
	Clicks:	[4]	1	3	[9]	1	2	2	[2]	0	1	
Fail %		20%	60%	0%	40%	0%	0%	20%	20%	0%	40%	
Novice Average	Seconds:	147	112,4	74	146,8	32,4	71,6	58,8	29,4	12	41,4	725,8
	Clicks:	3,8	3,8	2,8	4,6	1,8	2,8	2,8	2	0	1,6	26
[ ] = Fail												

Figure 21 – Table showing the statistical data from the Think-Aloud test session with novice participants. If a result is marked with red, it means that the result is worse than the same results from the experts. Yellow means that the result is the same, and green means that they did better than the experts.

EXPERTS												
		Task 01	Task 02	Task 03	Task 04	Task 05	Task 07	Task 08	Task 09	Task 10	Task 11	Combined
[E1]	Seconds:	70	102	75	[136]	27	30	35	40	[40]	23	
	Clicks:	2	5	4	[4]	2	3	2	3	[0]	2	
[E2]	Seconds:	100	[102]	80	20	18	28	23	11	X	10	
	Clicks:	4	[6]	4	2	1	2	3	2	X	1	
[E3]	Seconds:	70	50	15	26	7	29	26	4	10	26	
	Clicks:	5	4	1	2	1	3	2	1	1	2	
[E4]	Seconds:	227	187	40	53	7	89	[150]	48	11	[90]	
	Clicks:	5	3	1	3	1	2	[3]	2	0	[1]	
[E5]	Seconds:	[240]	92	12	[255]	X	47	56	19	30	10	
	Clicks:	[5]	5	1	[11]	X	2	3	2	0	1	
Fail %		20%	20%	0%	40%	20%	0%	20%	0%	40%	20%	
Expert Average	Seconds:	141,4	106,6	44,4	98	14,8	44,6	58	24,4	22,8	31,8	586,8
	Clicks:	4,2	4,6	2,2	4,4	1,3	2,4	2,6	2	0,3	1,4	25,4
[ ] = Fail												

Figure 22 - Table showing the statistical data from the Think-Aloud test session with expert participants.

4.12 Observational Think-Aloud Data

In this section, we will present the findings and observations from the approximately five hours of Think-Aloud videos (The complete set of notes and videos from the Think-Aloud can be seen in appendix 6). The content that we will present in this section is split into 13 parts (the 13 tasks the participants were asked to complete). Each task part contains the task description for that task, the observation notes, and the times spent and number of clicks used by the participants. As mentioned in the research design section, in the first part of the Think-Aloud research design, we asked the participants to find content that was not mountain bike specific content. This was for example "How do you book a stay at Feriecenter Slettestrand?" and "Where can you see the prices for a week stay at Feriecenter Slettestrand?" In the last part of the Think-Aloud research design, we only asked the participants to solve mountain bike specific tasks.

Below we will present the results from the Think-Aloud tests. Tasks 1-9 are completed on Feriecenter Slettestrand's main website and tasks 10-12 are completed on the mountain biking website. The last task, number 13, is referring to both websites. In each of the first 12 tasks the first decision/click the participants made is presented. In some tasks the next one to three decisions/clicks has been mentioned due to fast browsing between pages. This is done to highlight the first impulse from the participants, which will give us an overview of their decisions for our analysis. It is interesting to see how the participants react to the tasks and what decisions they make, when comparing the two groups. This section will however not focus on or in any way analyse the similarities or differences between the two groups. This will be done in the analysis section. This is done to make sure that this section only presents data from the Think-Aloud and not to further examine that data. To further present the results from the videos, the average time used and clicks made by the novices [N1, N2, N3, N4, N5] is mentioned in each of the tasks (extracted from figure 21 and 22). The same is done with the time used and clicks made by the experts [E1, E2, E3, E4, E5]. Each task is introduced with the task description given to the participants. It is important to acknowledge that the results are from the approximately five hours of video material collected during the Think-Aloud testing and does therefore not contain every decision taken by the participants nor their statements. The full Think-Aloud videos can be found in appendix 09.

Task 1

Task description: “Find a suitable holiday home for four adults and two children”

One of the novices looks at “Priser” [N1], when trying to find a price for a suitable apartment for four adults and two, while one looks at “Ferieboliger” [N2] and three click on “Ophold” [N3, N4, N5].

Three of the experts click on “Ophold” [E2, E3, E5], while one click on “Booking” [E4] and one click on “Ferieboliger” [E1].

Average time used and average clicks made in task 1:

TASK 1	**Clicks**	**Seconds**
Novices	3,8	147
Experts	4,2	141,4

Task 2

Task description: “Find the opening hours for the reception.”

Four of the novices click on “Kontakt” [N1, N2, N4, N5] and one novice clicks on “Ophold” [N3]. The information they are looking for is located in “Kontakt”, but they do not see it, as the information is concealed within a text field. The novices then click on “Om os”, “Værd at vide” and “Forespørgsel”.

Two of experts click on ”Kontakt” [E1, E2], one clicks on ”Kursuscenter” [E3], one click on ”Værd at vide” [E4] and the last expert click on ”Om os” [E5].

Participants from both groups also click on “Booking” and “Ophold”, as they say it would make sense to them that something that has to do with checking in to the holiday centre, would be located in categories for booking a vacation.

Average time used and average clicks made in task 2:

TASK 2	**Clicks**	**Seconds**
Novices	3,8	112,4
Experts	4,6	106,6

Task 3

Task description: “Find three activities for your stay at Feriecenter Slettestrand, whereas one activity is for children, one is for adults and the last is a task for the whole family”.

Four of the novices and four of the Experts click on the menu “Aktiviteter” [N1, N3, N4, N5, E1, E3, E4, E5], while one of the novices click on “Aktiv i Naturen” [N2] and one of the experts click on “Mountain bike” [E2].

Average time used and average clicks made in task 3:

TASK 3	**Clicks**	**Seconds**
Novices	2,8	74
Experts	2,2	44,4

Task 4

Task description: “Find the price for an included demi pension (breakfast and lunch) with your holiday stay.”

Three novices click on “Ophold” [N1, E3, E5], while the last two novices click on “Restaurant” [N2, N4].

Two experts click on “Ophold [E1, E5] and one clicks on “Priser” [E4], while the two remaining experts click on “Restaurant” [E2, E3]. In this task we then see two experts and two novices click on the same label; “Restaurant”, which is where the price information for including a demi pension is located.

Average time used and average clicks made in task 4:

TASK 4	**Clicks**	**Seconds**
Novices	4,6	146,8
Experts	4,4	98

Task 5

Task description: “Where can you book your vacation at Feriecenter Slettestrand?”

Three of the novices click on “Booking” [N1, N2, N5], while one click on “Ophold” [N3] and the last clicks on “Ferieboliger” [N4].

Four of the experts click on “Booking” [E2, E3, E4, E5], while the last expert clicks on “Ferieboliger” [E1].

Average time used and average clicks made in task 5:

TASK 5	**Clicks**	**Seconds**
Novices	1,8	32,4
Experts	1,3	14,8

Halfway through the tasks the theme of the tasks is directed towards mountain biking and the mountain biking related content of Feriecenter Slettestrand’s two websites.

Task 6

Task description: “What mountain bike specific content do you expect to discover? Name three things.”

This was done in order to see, what the novices and experts thought to be important, when booking a mountain bike themed vacation at Feriecenter Slettestrand. The answers from the participant start to vary, but some of them are still answering the same. Three novices [N2, N3, N4] and two experts [E2, E5] want to be able to find information on bike renting. Four out of the five novices want to know something about the tracks [N1, N3, N4, N5].

One of the novices [N2] mentions that he does not care about the tracks (One of the novices [P20], who only participated in the Card Sorting, mentioned that “Sporenes tilstand” is redundant, as he was sure that they were always perfect, as he did not think that Feriecenter Slettestrand would write about bad tracks).

All five experts mention that they want to read about the tracks [E1, E2, E3, E4, E5], while two of them [E3, E5] want to read in further detail about the present condition of the tracks. One of them [E3] highlights that it is important that the information on present track conditions are updated regularly.

One of the novices [N3] struggles to come up with things the participant hopes to see on the website in relevance to mountain biking and only mentions two things (instead of three). Another novice [N1] comes up with three, but mentions that due to lacking knowledge on the mountain biking subject, he is not sure if the information he seeks is important.

Task 7

Task description: "Find out how long the Svinkløv Klitplantage track is."

All five novices and four of the experts click on "Mountainbike" and then further click on "Stier og Ruter" [N1, N2, N3, N4, N5, E1, E3, E4, E5]. Only one of the participants takes another direction. One of the experts [E2] uses the footer navigation and clicks on "Stier og Spor". This means that this participant does not need to click on "Mountainbike".

Average time used and average clicks made in task 7:

TASK 7	**Clicks**	**Seconds**
Novices	2,8	71,6
Experts	2,4	44,6

Task 8

Task description: "Find the course "Trailbuilder kursus" held by Feriecenter Slettestrand."

Four of the five novices click on "Mountainbike" [N2, N3, N4, N5] and one novice clicks on "Kursuscenter" [N1]. The same happens for the experts, where four experts click on "Mountainbike" [E1, E2, E3, E5] and one expert clicks on "Kursuscenter" [E4].

Average time used and average clicks made in task 8:

TASK 8	**Clicks**	**Seconds**
Novices	2,8	58,8
Experts	2,6	58

Task 9

Task description: “Find the date for a race held by Feriecenter Slettestrand called “Slettestrand Ultra” (This is a race that takes place in October 2015.)

Three novices click on “MTB Events” [N1, N4, N5], while two novices click on “MTB Aktiviteter” [N2, N3].

Two of the experts click on “MTB Event” [E2, E5] and one expert clicks on “MTB Kalender” [E4], while two experts click on “MTB Aktiviteter” [E1, E3].

Average time used and average clicks made in task 9:

TASK 9	**Clicks**	**Seconds**
Novices	2	29,6
Experts	2	24,4

Task 10

The participants are now introduced to the mountain bike specific website by Feriecenter Slettestrand.

Task description: “Locate a possible method for receiving mountain biking relevant news.”

All five of the novices [N1, N2, N3, N4, N5] tells us that they would use the newsletter functionality located on the bottom left side of the website and does not for example mention “Event” from the left side menu.

Three of the experts click on “Events” [E1, E2, E3], but two of them [E1, E2] had hoped for a newsfeed element located on the front page. One of the experts [E4] would use the newsletter functionality and the last expert [E5] would just login to Facebook and find Feriecenter Slettestrand’s Facebook page, which he expects will be updated more regularly than the mountain biking website.

Average time used and average clicks made in task 10:

TASK 10	**Clicks**	**Seconds**
Novices	0	12
Experts	0,3	22,8

Task 11

Task description: “Find a method for booking a mountain bike specific vacation in Feriecenter Slettestrand.”

Four of the novices [N1, N2, N3, N4] click on “MTB Ferie”, but only two of them [E3, E4] find the relevant link, which sends the participant to Feriecenter Slettestrand’s main website, which is where users should go for booking vacations. One of the novices [N1] then clicks on “Kursus”, while another [N2] clicks on “MTB Camp”. The fifth novice [N5] clicks on “Ophold” in the middle of the front page and mentions that he finds it difficult to differentiate between “Ophold” and “Ferie”.

Four of the experts [E1, E2, E3, E4] click on “MTB Ferie”, but one [E3] tells us that he does not find relevant information, so he explains that he would rather click on the big Feriecenter Slettestrand link-element on the right side of the website, which is a link directly to the main website. The last expert [E5] also clicks on the Feriecenter Slettestrand link-element.

Average time used and average clicks made in task 11:

TASK 11	**Clicks**	**Seconds**
Novices	1,6	41,4
Experts	1,4	31,8

Task 12

Task description: “Where can you find pictures relevant to Feriecenter Slettestrand and mountain biking?”

Three of the novices [N1, N3, N4] click on “Blog”, while another [N2] clicks on “MTB Ferie” and “Sporene”. The last novice [N5] clicks on “Oplev Eventyret”, which is a link to a blog post. [N3] tells us that it would make sense with a gallery menu for pictures.

When looking at the experts, one expert [E1] clicks on “MTB Ferie” and then “Sporene”. Another one of the experts [E2] finds an Instagram feed with pictures in the bottom of the front page. The third expert [N3] clicks on “Om os” and “Teamet”. The last expert [E5] clicks on “MTB Kurser”, “Om os” and “MTB Camp”. Two of the experts [E2, E4] do however mention that there could be a label in the menu called “Billeder”.

Task 13

In task 13 the participants are asked to give their estimate on their overall satisfaction of the two websites and how the websites work in collaboration. First the statements for the main website are presented and then the statements for the mountain biking website are presented. The statements have been categorised into labels, functionality, navigation, visual output and other remarks. First the statements from the novices are presented in each category and then the statements from the experts are presented in each category. If there are no statements presented from neither the novices nor experts in a given category, it means that they did not contribute with any statements in that given category.

4.13 Summary of the 13 Tasks

In this section we presented the average amount of clicks and time used by the participants in the Think-Aloud tests, combined with a description of the first clicks made by the participants. This presentation will be useful for us in the analysis section, where this data will be valuable for us in comparing the differences and similarities between the two user groups.

4.14 Comments from the 13 Tasks

In the next section, we will present the commentaries made by both user groups during the Think-Aloud test. The commentaries will be categorised into the groups: labels, functionality, navigation and visual output. Each group will be separated into two groups: novices and experts, so that the commentaries from the two groups are easily distinguished (some of the categories will only have commentaries from one of the groups, as the other group did not say anything in regard to the categories). First the comments for the main website will be presented, which is followed by comments for the mountain biking website. The four groups: labels, functionality, navigation and visual output have been chosen, as they are important in regards to usability and Information Architecture. The content of the comments were also important to the users and what they expected of the two websites, which is also interesting in regard to our hypothesis, as we want to find out what the two user groups looks for in functionality. The comments will be presented with the number of participants mentioning the same thing (for example [N1, N3, N4] mention that they find the amount of text on the main website for excessive). This is presented in bullet form as to clearly separate the different issues mentioned by the participants.

Main Website

Labels

Novices

- [N1] talks about the labels and how they should be examined and mentions that he does not read the text on the front page.
- [N1, N4] mention that it is difficult to understand and differentiate the labels “Events”, “Kurser”, “Aktiviteter”, “Ferieboliger and “Ophold”.
- [N5] mentions that there is a lot of information on the main website in a negative fashion.

Experts

- [E2] does not see the difference between “Event” and “Aktivitet”. Does “Undervisning” for example fit under “Event” or “Aktiviteter”?
- [E1] likes the “Om os” menu, as he sees it as honest.
- [E2] some labels confuse him.

Functionality

Novices

- [N1, N2, N3] want a booking functionality.
- [N2] needs “Aktiviteter”, “Priser” and “Booking” to be more in prominent on the main website.
- [N4, N5] need the labels in the middle of the front page to be interactive, so that visitors can click on “Aktiv i Naturen”, “Ophold”, “Kursus” and “Restaurant”, which are not interactive as of now.

Experts

- [E1, E3] want relevant newsfeed functionality, where the tracks condition is highlighted.
- [E1, E2] notice that the navigation scheme is different on the mountain biking website, which they find to be negative. The connection between the main website and the mountain biking website is lost. The visual output is also different between the two websites.
- [E3] thinks the site needs a booking functionality.
- [E5] needs a search functionality on the main website.

Navigation

Novices

- [N2, N3, N4, N5] think that there are too many possibilities in regards to navigation (two top navigational menus and one navigational menu in the footer) and needs more focus on getting a better overview. It is difficult for them to get an actual overview of the main website and understand how the navigation is meant to be used.
    - [N5] likes the menu in the footer better than the two top menus.
    - [N5] would like a drop-down menu.

Experts

- [E2] would delete the menu in the footer, as it shows the same as the top navigation.
- [E2, E5] mentions that it is difficult to get an overview of the content, as there is a lot of content. They like the content; there is just too many menus that look alike, which confuses them.

Visual output

Novices

- [N2, N4] want more focus on mountain biking on the main website, as they regard Feriecenter Slettestrand as one of the best places for mountain biking. There are no visual clues on the front page, telling visitors that mountain biking is a high priority.
- [N3] The red font colour used for the introduction text annoys the participant, as the participant is used to reading black text on websites.
- [N3] needs more pictures from the area and the Feriecenter Slettestrand holiday centre.

Mountain Biking Website

Labels

Novices

- [N1] does not know what "Sporene" refers to, as the participant sees it as indefinable.
- [N2] likes the visual elements (background picture/visual feel) of the mountain biking website.

- [N5] needs a bigger focus on “MTB Ferie”, “Events” and “Kurser” on the mountain biking website.

Functionality

Novices

- [N2, N3] like that it is possible to click on a link and then be sent from the mountain biking website over to the main website.
- [N3] needs are more eye catching booking functionality on the mountain biking website.

Experts

- [E4] needs there to be a focus on the tracks on the mountain biking website (the condition and description of the tracks).
- [E4] sees it as important that visitors can rent bikes directly on the mountain biking website.
- [E4] wants a focus on the Feriecenter Slettestrand’s restaurant and their other facilities.

Navigation

Experts

- [E3] uses Facebook to keep track on updates and wants there to be a Facebook link or feed on the mountain biking website.

Visual output

Novices

- [N2] finds the picture chosen for the front page (which is a picture linking to a blog post), to be too large for the front page, as it fills the entire view. It does not lead to the content below the picture (many of the participants did not look at the middle or bottom of the front page).
- [N4] needs a bigger focus on the mountain biking website telling that the website is also a part of a holiday centre.

Experts

- [E1, E5] find the picture chosen for the front page (which is a picture linking to a blog post), to be too large for the front page and it is difficult to get an overview of the page as “too much is happening”.
    - [E2] likes the white background from the main website better than the pictures used on the mountain biking website. [E2] mentions that he does not like pictures as backgrounds. [E5] likes the visual output of the main website.
    - [E2] mentions that too much important content is hidden in the bottom of the front page.
    - [E2] mentions that the content field is too narrow compared to how much space is “wasted” around the content (in the right side).
- [E4] likes the “more fluent” visual output of the mountain biking website than compared to the main website.

The comments in the four categories above will be used in the analysis when they are relevant. The introduction to the analysis section will go into detail in how we use the data from the data presentation.

5.0 Analysis: Introduction

Feriecenter Slettestrand's websites have different target groups. The two websites we have focused on are their main website, which has a dynamic range of at least four different target groups, and their new mountain biking website which primarily is aimed at one specific target group within the domain of mountain biking. Since it is nearly impossible to narrow the target group of the main website down to a single target group, we have used it to contrast the difference in evaluating the usability of websites with a non-specific domain and target group, and websites with a specific domain and target group.

The reason for doing so lies with the idea that there is not always a *one-fits-all* solution when evaluating usability and Information Architecture (as enlarged on in section 2.6). Much of the literature concerning these areas have conflicting accounts on *when* and in *which* contexts participant domain expertise plays an important role, and why. In earlier sections, we have summed up the various views of the matter, which might be the results of a shifting focus in the web usability discipline. In his 2000 article, usability expert Jakob Nielsen proposed his explanation to why there is such difference of opinion, and the reasons for why the focus has shifted so much in the last few decades, arguing how it was very much dependent on who you were trying to design for, rather than what you tried to design (Nielsen, 2000). Why would you use domain novices for evaluating the usability of an interactive website or interface that was targeted at experts within a very specific domain?

Other studies and researchers, (for example Bednarik & Tukiainen, 2005; Botella et al., 2014; Dou et al., 2009; Karapanos et al., 2008; Kinney et al., 2008; Kjeldskov et al., 2010; Lazonder et al., 2000; Nielsen & Molich, 1990a; Nielsen, 1992; Nielsen, 2000) have dwelled into this problem, trying to expound on the significance of domain expertise variables within various contexts, and there is, seemingly, still no definitive answer that can tell you exactly when and in which contexts it matters most. However, it seems that there is an agreement that domain expertise influences the way that users utilise and perceive systems, such as websites, and that the way users behave can influence evaluation processes (Benyon, 2010, p. 32; Wills & Hurley, 2012). The question is rather how evaluation test participants' domain expertise (or the lack thereof) influences the results of the evaluation, and if there are situations in which these influenced results can be used to heighten the quality of the overall evaluation, if they were expected and included in the design of the evaluation process.

This disagreement within the usability evaluation discipline lead us to our hypothesis, in which we aimed to investigate how the results of a usability evaluation are influenced by domain expertise and evaluation contexts. The focus for this thesis has not entirely been with the aim of completing a usability evaluation of Feriecenter Slettestrand's websites, but rather to also investigate where two user groups' (one with a defined domain expertise, one without) usability evaluation results differentiate, in order for us to extract the contexts of where these differences exists. The goal of doing so, is to be able to distinguish between the evaluation contexts where participant domain expertise matters more than others, and use this distinguishing to contribute to the usability and Information Architecture evaluation discipline, in an effort to help future researchers decide on when and where to use domain experts as participants for their evaluation tests.

For us to be able to use the results from the evaluations we completed of the two websites and with the two participant groups, we will divide the analysis into the component parts (or attributes) of usability and Information Architecture disciplines. This means that we will investigate each part of the evaluation by itself (for example *learnability*, *navigation* and *labelling*), both for the sake of evaluating the websites, but also (if not more) to answer our hypothesis concerning participant domain expertise, and how it influences the evaluation process and results.

Each of these components are represented by one or more of the tasks that the participants were asked to solve during the evaluation process (primarily the Card Sorting and Think-Aloud tests), and the tasks are representative of how typical users would, in most cases, use and utilise the websites for solving real life tasks or complete goals. Since all the participants, regardless of domain expertise, solved the same tasks in the evaluation process, we can compare the results of the two participant groups, and determine in which circumstances we find either participant group to serve its purpose in the best possible way, or in which contexts either participant group lacks or overlooks important elements of the evaluation. We will also include some of the results from the Heuristic Evaluation, even though the results from this test method are based on expert assessments and accepted, relevant theory rather than actual user testing. Nevertheless, the results from the Heuristic Evaluation are still relevant, and can be used in combination with the results of the other tests to evaluate and relate the usability issues to specific design components.

After having gone through the adequate number of usability and Information Architecture attributes in the analysis, and summarised our findings for each of these attributes, we can combine the findings into a discussion of when we found the influence of test participant domain expertise to be crucial to

the evaluation process, and in which contexts and situations it is less crucial when evaluating. The actual evaluation of the websites is of course unavoidable, and will be concerned too, but should be considered the second goal of the analysis.

To sum up, the overall idea of the analysis is to acknowledge that the term *usability* (and to some degree also *Information Architecture*) are terms that can change their meaning depending on how the definition is formed, and in which context it is being used (we enlarged on that in section 2.1). We find it important to acknowledge this, because, depending on what you are evaluating, you have to take into consideration the methods and approach you will be using to make sure that the evaluation process fits the goal of the evaluation. Therefore, we also find it important to take into consideration the domain expertise of the participants used for the evaluation process, as this is also a variable that is agreed can influence the data derived from the tests. We do not think that there is an *always-right* answer as to when to use domain experts or novices, but that it is dependent on the goal and focus of the evaluation. Or as Morville and Rosenfeld describe it; "*No single approach can stand alone as the one right way to learn about users and their needs, priorities, mental models, and information-seeking behavior. This is a multidimensional puzzle—you've got to look at it from many different perspectives to get a good sense of the whole*." (Morville & Rosenfeld, 2006, p. 247).

We have already given an example of the definition of usability that we have used earlier (section 2.1), and will continue using, but understand that it might be different from how other people use it. This is also one of the reasons for why we split the analysis into parts consisting of the attributes from our definition, so that each attribute can easily be connected to other, similar or dissimilar, definitions of usability using the same or similar attributes. As a consequence, this last analysis section should not be considered one, large "usability evaluation", but rather a combination of attributes that all relate to usability evaluations in general, in an effort to contribute to the discussion of when it is the best practice to involve domain experts or novices as participants in usability and Information Architecture evaluations.

5.1 Usefulness, Effectiveness and Efficiency

In this section, the analysis will focus on three usability attributes: *usefulness*, *effectiveness*, and *efficiency*. We will first present the three aspects (which we also refer to in theory section 2.1). Subsequently we will point out relevant encounters where the two participant groups found it easy or difficult to reach their goals (how *useful* they found the websites to be). Then we will look at the clicks made and time spent by the two participant groups in relevance to evaluating the *effectiveness* and *efficiency* of Feriecenter Slettestrand's websites.

The *usefulness* refers to how usable the website is and if the users can succeed with their goals in a useful manner (Rubin & Chisnell, 2008, p. 4). This is both in regard to how the website enable users and if the website gives users a method for completing their goals; *"The user can do what he or she wants to do the way he or she expects to be able to do it, without hindrance, hesitation, or questions*." (Rubin & Chisnell, 2008, p. 4).

The *effectiveness* of a website shows how the website act to a degree, which the users find satisfying, for example does the website do what the users expect it to do (Rubin & Chisnell, 2008, p. 4), or does the website let its users find the information or functionality they are looking for when clicking on the websites labels. This can be measured by looking at the clicks that the participants make in a Think-Aloud test (figure 21 and 22). It is relevant to look at where the participants click (which labels do the click on) and how many clicks they make, as to see how effective a website is in guiding its users to the correct information. If the participants expect to find a price list, when clicking on a label called "Prices", then that information should be available.

In regard to *efficiency* we look at how the participants can use the system after the learning curve has ceased to rise, which can be measured in time (Nielsen, 1993, pp. 30-31; Rubin & Chisnell, 2008, p. 4). The participants were not introduced to the two website before the Think-Aloud tests took place, but they were all using the websites for 20-30 minutes during the test and did therefore get a chance to somewhat learn how to use the website (The learnability of the two websites by the two participant groups will be analysed in section 5.6). We will refer to the time spent in figure 21 and 22, when analysing the *efficiency*. The time will tell us how fast the participant groups completed the tasks, which can show us if the websites are easily navigated and enables the users to find relevant information within appropriate time.

5.2 Evaluating Usefulness through Participant Comments

To evaluate the *usefulness* of Feriecenter Slettestrand's two websites, we must look at what the users can do on the given websites and what the users actually expect to be able to do in in regard to functionality. To see if the two participant groups experienced that Feriecenter Slettestrand's websites complied with what the participants experienced as *usefulness*, we will look at the data from the Think-Aloud test. In the Think-Aloud test we tasked the participants with finding various information (like how to book a stay at Feriecenter Slettestrand), which opened up for participants to give their opinion on various functionalities. In the last tasks we also asked the participants if they were missing anything (functionality, information, content etc.), which also gave the participants an opportunity to speak their minds. This also gave us information regarding usefulness. As we are evaluating websites, where a large amount of the content is mountain biking content and a big part of the visitors at Feriecenter Slettestrand are mountain bikers, we asked the participants to list three things that they would expect to find information on in regard to mountain biking. When conducting evaluations it is a relevant task to give the participants, to acknowledge the participants' expectations. When doing evaluations the literature tells you to find participants who are potential users and have domain knowledge, which is the common method for doing evaluations and is what we want to examine further (Jenkins et al., 2003; Rubin & Chisnell, 2008, p. 115; Sova & Nielsen, 2003, pp. 29-31). It is therefore interesting to see what the two participant groups that we have assembled expected to find on a website mainly for mountain bikers. In the first subsection, we will present the goals expected by the experts and then in the next subsection we will present the goals expected by the novices.

Goals Expected by the Experts

Two of the five domain expert participants [E1, E3] expressed a disappointment in regards to not finding a calendar functionality or news feed on the main website. They expected a website like Feriecenter Slettestrand's main website to contain some form of updates to what is happening in the near future, for example information regarding an upcoming mountain biking race, as expert [E3] was looking for or information on the condition of the mountain biking tracks, as expert [E1] was looking for. For the mountain biking website, expert [E4] would expect regular updates on the condition of the tracks to be a main focus point for the website. This is also a feature mentioned by expert [E3], which he thinks can bring value to the mountain biking website, as he sees this as one of the main goals for the visitors of the mountain biking website.

Goals Expected by the Novices

Two of the domain novice participants [N3, N4] mention, when asked what they were missing on the main website, that they expected there to be a larger focus on the visual representation of Feriecenter Slettestrand. Novice [N3] mentions that she would find it to be relevant if there were a gallery menu, and novices [N2, N4] mention that the main website does not visually show that Feriecenter Slettestrand has a focus on mountain biking, which is strange as Feriecenter Slettestrand "*is the place for mountain biking in Northern Jutland*"[22] (cited from [N2] appendix 09, video 08 (27 minute mark))[23]. As of now, Feriecenter Slettestrand's main website does not contain any gallery and does not give its visitors any direct means to visually experience the centre or its surrounding environment.

Similarities in Goals Expected by both Participant Groups

When we asked the participants to find out how to book a stay at Feriecenter Slettestrand, they all experienced that the websites do not have an online booking functionality, which nine of the ten participants [N1, N2, N3, N4, N5, E1, E1, E3 and E4] told us they found to be disappointing, as this is a functionality that they would expect to be available on a website like Feriecenter Slettestrand's main website. It was especially disappointing when the main menu in Feriecenter Slettestrand's main website contains a label called "Booking" (as seen in figure 23), which gave the participants the impression that online booking was available.

Figure 23 - Here the "Booking" label can be seen in the top menu (third from left).

When looking at the goals expected by both user groups we see some similarities in regards to both mountain biking specific content and in regards to common goals like online booking. Both groups are highly interested in the possibility of online booking and are disappointed when this is not available. The suggestion for Feriecenter Slettestrand is therefore the same, even if we would only have used one of the two participant groups for our evaluation, as both groups expect to be able to book a vacation online. This can be seen as a common user goal that users, disregarding their domain expertise, would expect to be able to complete when visiting a website for a holiday centre.

[22] Quotes from the Think-Aloud data are being translated from Danish to English by the authors.
[23] This quote, and the following quotes from participants can be found in the videos from the Think-Aloud test, all in appendix 09.

If we then look at a goal like being able to find information on upcoming events or mountain bike races, which could be part of a calendar or news feed, it is only the domain experts who mention this goal in regard to mountain biking and who expected this to be an available functionality on the main website. The novices do not mention this functionality in regard to mountain biking, however one novice [N3] does mention that she would like to see a week program, so that she could plan activities ahead of her stay. This is however not in relation to mountain biking, but more a detailed list of the various family activities Feriecenter Slettestrand offers (like horseback rides, pancake making, stone grinding etc.). The suggestion for Feriecenter Slettestrand in an evaluation would then be two different suggestions, if we only used one of the participant groups for a Think-Aloud test. The calendar availability would not be relevant if we only used the domain novice group. A valuable functionality for the experts would therefore be lost, if we had only used a novice user participant group.

What is surprising is however that it is two of the novices that express a disappointment in regard to the visual aspect of Feriecenter Slettestrand's main website. The two novices mention that the main website does not visually inform them that they have landed on a website for a holiday centre with a large focus on mountain biking. As seen in figure 24, the front page of Feriecenter Slettestrand's main website does no visually tell its visitors that they are visiting a website where mountain biking has a large focus. This is possible because of the fact that the experts participants had all been to Feriecenter Slettestrand before and used their track and through that already know how the centre and its surroundings look like. They therefore already know what it looks like and so the thought of seeing it visually is not something they mention.

When looking at mountain biking specific content, the novices [N1, N3, N4, N5] and experts [E2, E3, E4] agree that track-information and maps is a type of information they expect to find. Therefore, finding this type of information could be one of the usability evaluation goals, as there is a common interest in the availability of this information. The same goes for the common interest in renting bikes, as two of the experts [E2, E5] and three of the novices [N2, N3, N4] expect to be able to rent mountain bikes at Feriecenter Slettestrand, thus renting bikes could also be used as a usability evaluation goal, as both of these goals are not limited to one participant group with a specific domain expertise.

Both groups also have one participant each [N1] and [E4] who mentions that it would be useful to see difficulty ratings on the various tracks. Two novices [N4, N5] and two experts [E1, E3] also express interest in mountain biking activities that they can participate in when visiting Feriecenter

Slettestrand. One expert [E4] and one novice [N1] also both ask for information on where they can store their bikes safely when staying at Feriecenter Slettestrand. Expert participant [E4] further expresses that he needs to be able to clean and repair his bike, if that would become necessary. The experts [E2, E5] also mention that they would like to know what mountain biking relevant facilities Feriecenter Slettestrand have. None of the novices mention this, showing that they do not either find this interesting or they do not think about this when asked. When going into details like this it is clear that the novices do not think about aspects like fixing bikes and what other facilities Feriecenter Slettestrand have.

Figure 24 - This is the first view the visitors at Feriecenter Slettestrand's main website see and it does not, as mentioned by two novices, visually present the fact that Feriecenter Slettestrand have a large focus on mountain biking.

As seen in this section, there are similarities between the two groups, when they are asked what they expect the main website to enable them to do in regard to useful goals and user goals. There are however also various goals that the novices do not mention and one of the five novices also expresses that he does not have the appropriate knowledge on the subject to provide, what he thinks, is the foundation for a good answer to the question "What three things do you expect to find in regard to mountain biking?" Another novice could only provide two suggestions to the question, which would make the Think-Aloud results not as thorough for the group of domain novice participants, as it would for the group of domain experts.

There are, however, many similarities between the answers from the two groups as we have highlighted in this section. If we look at the time requirement necessary for the planning of Think-Aloud tests with domain novice participants, compared to using only domain expert participants, there is a huge difference in the time required for recruiting participants for the Think-Aloud, making the planning of a time schedule for the test process, finding locations where it is possible to set up controlled test setting, and the time used for travelling, very time-consuming.

The novices are "random" or people we know, who would like to participate in a Think-Aloud. Those people can be taken into a test without a large amount of time or effort, in comparison to the experts, whom we used online mountain biking forums and connections at Feriecenter Slettestrand to come in contact with. When analysing the Think-Aloud results, where participants are asked to list useful information in regard to mountain biking, the novices are close to listing the same information that the experts list, indicating that the novices' contributions are nearly as valuable in usability evaluations.

5.3 Evaluating Effectiveness through Clicks

In this section we will look at how *effective* the two participant groups were in locating the various *useful* goals highlighted in the section above. These goals are useful to the participants and the two websites should therefore help the participants finding those functionalities *effectively*. We will use task 5 (*Find out how you can book a stay at Feriecenter Slettestrand*) as an example for this part of the analysis, as the results were valuable, because it was a task that both participant groups expressed concern with. When we asked the participants to find out how to book a stay at Feriecenter Slettestrand, three novices [N1, N2, N5] and four experts [E2, E3, E4, E5] clicked on "Booking". After clicking they found that the website does not contain an actual online booking functionality, but instead the users were presented with a telephone number and an e-mail. Both user groups expressed that they were disappointed with this and that they expected a booking functionality to be available, as this is what they expected when seeing the label "Booking". However, one expert [E5] does mention that he is fine by this, as he would call them anyways to book a vacation, but he can see that it would be bothersome for other users, not to have an actual online booking functionality. One novice [N4] and one expert [E1] clicked on "Ferieboliger" as their first click for this task, and the last novice [N3] clicked on "Ophold" as her first click, as all three expected to be able to choose a vacation home and from there be navigated to a booking functionality. It is therefore clear that the participants (from both domain groups) expected to be able to complete the goal of booking a vacation online and

that Feriecenter Slettestrand in order to satisfy its potential guests, should implement an online booking functionality that can be accessed either through a navigational label called “Booking” or through the pages “Ferieboliger” and “Ophold”, so that users can complete the goal effectively. This shows that there would be no difference in the suggestion, even though we only used the novices.

When considering the average amount of clicks made by the two groups when completing task 5, the novices made 1,8 clicks and the experts made 1,3 clicks in average, making that a difference of 32%. It is however only one novice [N3] that makes the difference, as novice [N3] used four clicks, compared to the rest making one to two clicks, which made the average clicks go up marginally. 100% of the participants do however find a way to locate the information regarding booking and within a few clicks. The *effectiveness* of locating a functionality like booking, which is *useful* to the both participant groups is however not as efficient as it should be, which will be elaborated in the next subsection.

When given a task like finding a booking functionality both domain groups mainly clicked on “Booking”, with two deviations in the novice group and one deviation in the expert group. One of the deviations, a participant clicking on “Ferieboliger”, happened in both user groups. There is therefore nothing that differentiates the two user groups apart in this task, as the behaviour in both user groups were close to identical, which would lead to the same suggestion in a usability evaluation.

When looking at the overall average clicks made by the two participant groups in all the tasks there is a difference, as the experts used 0,6 less clicks throughout the given tasks (figure 21 and 22). This is not a large margin and shows that for experts and users on a website like Feriecenter Slettestrand’s, the difference in being a potential user with domain specific knowledge does not have a real impact when evaluating the effectiveness of a website like Feriecenter Slettestrand’s websites. There is however a relative large difference in the time used by the two participant groups, which we will analyse in the next section.

5.4 Evaluating Efficiency through Time

In this section, we will look at the *efficiency*, in which the two groups found useful information on booking. The average time for locating the “Booking” label and finding the relevant information notifying them that they have to make a call or send an e-mail was for the novices 32,4 seconds and 14,8 seconds for the experts, making that a 118,9% difference in time spent for that task. Novice [N3], who made the most clicks, was also the one using the most time on the task. She mentions after

her first two clicks, that she did not see the small white top-most menu, which includes the "Booking" label (figure 23).

As mentioned above, nine of the ten participants expected to be able to complete the goal of booking a vacation, which makes this a high priority goal. It should therefore be efficient for users to achieve this, which our Think-Aloud test data show is not the case. To make the website more efficient the *accuracy* (Rubin & Chisnell, 2008, p. 4) of finding the label "Booking" should make it possible to find it within seconds and with no prior clicks, as to improve the efficiency of the main website. A redesign proposal, suggested on the basis of our test results, is that Feriecenter Slettestrand should move the "Booking" label from the top-most menu down into the main menu, or implement an online booking functionality directly on the front page.

If we then look at a mountain biking relevant task like "Find out how long the Svinkløv Klitplantage track is", the novices used in average 27 seconds longer on that task than the experts did, which is interesting, as the novices only made 0,2 clicks more than the experts in average. All of the participants clicked on "Stier og spor", which shows that they all expect to find the relevant information there, but the novices used more time locating the actual information. This correlates with the theory by Russell-Rose and Tate, saying that double experts (participants with both relevant technical *and* domain expertise) use less time on each page reading the information (Russell-Rose & Tate, 2013, pp. 3-9). The reason for the experts to be faster than the novices is that double experts can "teleport" to the information they need. Experts [E2, E5] both use text search to make their search for information be more efficient, which also makes them faster than the rest of the participants in some tasks. This can then be seen in the average time spend by the two groups. We did not, when finding participants, recruit them based on their technical expertise, and the fact that two of the domain experts are also technical experts is a mere coincidence.

It is also important to notice that the novice participants spend more time on average examining the websites' content text, and that the novices read the text more carefully than the experts, the novices spend more time and thus making them less efficient, especially as the text on some content pages is overwhelming, as we will go further into in the next section.

Efficiency Lowered by Content Text and Font Colours

The efficiency of the main website is lacking partially due to the amount of text used to introduce each content page. Three novice participants [N1, N5, E5] complained about the large amount of text on the main website and most of its content pages (as seen in figure 26), which made the process of

finding relevant information on content pages slow, especially for the novices as mentioned above. The amount of text is also why two of the experts [E1, E2] uses text search to fasten up the process of finding the relevant information. They told us that they would not read the amount of text presented, simply because there is too much of it and it would take too long. Expert [E3] does not read the text, as he found it to be confusing. This diminishes the overall efficiency of the main website, as users will have to carefully read through a lot of text in order to find the information they need. To give an example of this, the second task in the Think-Aloud tests asked the users to find the opening hours for the reception (in case potential guests want to make sure that they can check in to Feriecenter Slettestrand within the opening hours of the reception). This information is located on the "Kontakt" page and four of the novices [N1, N2, N4, N5] and two of experts [E1, E2] actually click on "Kontakt", however the average time used is 112,4 seconds for the novices and 106,6 seconds for the experts and the average clicks are 3,8 for novices and 4,6 for experts, which shows a much higher number of clicks than expected, when six of the participants actually click on "Kontakt" as their first click. This shows that both user groups struggle to find important information, even though they are on the right page. As seen in figure 25, the information is concealed within the text and not highlighted. The participants do therefore not see the relevant information, making the main website less efficient to use.

Figure 25 - In this screenshot from the "Kontakt" page it is exemplified that important information can be hidden within text and that the use of the red and black font colours give the users a poor overview, which makes the website less efficient for the users.

Furthermore two of the novices [N3, N4] mention that they are annoyed with font colours, as actual text in the introduction is red and the embedded navigational links (Kalbach, 2007, pp. 92-93) in those introduction texts are black (as seen in figure 25 and figure 26). This confuses the participants, which may also distract them from the actual information. The two novices [N3, N4] mention that in their experience, the colours are reversed (black text and red links) and, as described by user experience designer James Kalbach, the most efficient method for embedding links in text, is to have the font blue and with an underline (Kalbach, 2007, p. 92).

Figure 26 - Here is an example of the excessive amount of text, which is making both user groups less efficient in finding relevant information. The font colour issue is also shown here, as the text is red and the embedded links are black.

5.5 Summary on Effectiveness and Efficiency

In the analysis of the usefulness, effectiveness and efficiency of Feriecenter Slettestrand's websites, we found both differences and similarities in the Think-Aloud test results from the two participant groups. Both groups wanted to be able to complete similar goals like booking a stay and finding information of the track, mountain bike rental, difficulty ratings and the storing of bikes. The differences between the two groups could be seen when it came to the functionality of a calendar that could tell when mountain biking races were planned. It was only the experts who mentioned this, which could be expected, as novices would probably not participate in races.

What was more surprising was that the experts did not mention the lack of pictures in regard to mountain biking. When it came to the visual experience of the website, it was only the novices that mentioned a lack of pictures showing that Feriecenter Slettestrand is a centre where mountain biking has a large focus. In regards to the study by Karapanos et al. (2008), where they studied the evaluation

of a smart TV remote by novice users and then four weeks later did the same evaluation, they saw that when the participants were novices, they focused on beauty and goodness the most (which include variables such as practical, manageable, presentable, innovative and simple). We can see the same here, as it was only the novices that mentioned the pictures, which can be categorised as presentable. The experts in the study by Karapanos et al. (2008) focused more on what they could do with the product and if it was useful for them in their everyday lives. This can be compared with our research, as the experts were more focused on the above mentioned calendar function, which the experts would interact with in order to find upcoming races.

In regard to effectiveness and efficiency we saw that the experts were faster than the novices in most tasks. The difference in clicks were however minimal and would not generate any large differences in a usability evaluation. The time difference was however more noticeable and the reason for the novices to use more time for each task can be explained through a few instances in the Think-Aloud tests; the novices complained about the amount of text and the use of font colours, which could be the reason for the experts to be faster. The novices used more time in general reading information and verbalising their thoughts on the website, than the experts. The results in general were however not that different, which correlates in some ways with the study by Lazonder et al. (2000), as they found in their study (which is also mentioned in the theory section) that domain knowledge was not the most significant factor, when searching for information on websites.

Overall the results we got from the two participant groups in our Think-Aloud test in regard to usability problems coming from usefulness, effectiveness and efficiency resemblances each other when comparing the two participant groups, as the two groups mentioned many of the same problems (goals that could not be completed, amount of text, font colours etc.). These results are comparable to the findings by Kjeldskov et al. (2010), who found that technical novices found an equal (and more) problems than the technical experts, even though their focus was on the technical expertise, rather than domain expertise. Novice participants, regardless if they are novices in terms of domain or technical aspects, can therefore still be useful in an evaluation context.

5.6 Learnability and Satisfaction

Knowing how to use a system or website, or rather how to use it most effectively in various circumstances, requires some amount of learning. Most often, learning *something* does not happen in an instant, but takes some amount of effort of the person(s) learning (Tullis & Albert, 2013, pp. 92-93). For learning to take place, it is essential that it is possible to gather experiences which can be manifested into knowledge of whatever you are trying to learn, which is called *learnability*. If there is no such way for experience to be usefully gathered, or the learning curve is too steep, the learning process can be very tedious, and the overall level of learnability is decreased and the usability is affected (Nielsen, 1993, pp. 27-30). Another usability attribute that is directly affected by the learnability, is (user) *satisfaction*. The satisfaction of a product refers to how the users perceive and feel about the product, metrics that can be quite difficult to measure, but the level of which can be indicated by observing how the product meets user needs, and how well the users adapt to the product and the ways it can be used. This includes how well the users learn to use the system and its features, making it more *usable*, as we described in the last section. To investigate this, we will primarily use data from the Think-Aloud tests.

5.7 Evaluating Learnability through Information Seeking and Navigation

To evaluate and measure the participant *satisfaction* of Feriecenter Slettestrand's websites, we must first look at some of the metrics of *learnability* found in our test data. We had no test tasks that inquired directly into the learnability of the two websites, but when surveying the Think-Aloud data, there are tendencies between the participants that could indicate how well they adapted to the website, one of them being the way the participants navigated and searched the websites for information that they were looking for. Since we will be focusing on *navigation* in another section, we will not focus on the way that the navigation has been built here, but rather how well the participants adopted the existing navigation.

Figure 27 – The main navigation of Feriecenter Slettestrand's main website. Notice how there is a smaller menu at the top also.

One of the things that struck us, when completing the Think-Aloud tests, was how many variants of navigation the participants used, and how they prioritised the possible options for navigation and information seeking.

Every participant, both experts and novices, used the top main navigation menu of the main website to navigate through the content and main pages of the website, almost every time they started a new task that involved finding a specific information. However, the top main navigation menu, as seen in figure 27 is a design feature that many of the participants, especially the novices, have a surprisingly hard time adapting to. The reason for this, seems to be that there are too many choices; the entire top main navigation (when including the smaller menu at the top of the other menu) has a total of 14 labels, where six of these are much harder to see because they are in the top-most menu. Four of the five novices [N2, N3, N4 and N5] mentioned during the Think-Aloud tests, that there are simply too many possibilities, and that they find it hard to get an overview of the navigation because of this. Two experts [E2 and E5] mention the same, and that they feel like the content is not the problem, but that there are too many menus with similar content on the website overall.

One of the issues with the menu manifests itself in the number of choices available, especially when taking into consideration the fact that in the top-most menu, "Home" is redundant (as you can click on their logo to go the their homepage) and "Booking", "Kontakt" and "Forespørgsel" are essential the same feature; there is no online booking, and the visitor is referred to the contact page ("Kontakt") instead. "Forespørgsel" is just an online formula the visitor can fill in, instead of writing an email himself, and could have been put under the "Kontakt" label as well.

When we look at the data we collected during the Heuristic Evaluation session (appendix 08), we did however note that the (top and bottom) menus are always visible and in the same location, no matter which page you are looking at, which is effective for global navigation (Kalbach, 2007, pp. 86-88; Morville & Rosenfeld, 2006, pp. 122-124; Tidwell, 2011, p. 80), as it can help users navigate and find their way around the content.

There is an issue with how the content pages on the main website are connected, as the top-menu are connected directly with the main pages of the site, but there are also many sub-pages that are only connected with each other in another local navigation sidebar menu (Kalbach, 2007, pp. 89-91), which appears under some of the main-pages, making the navigation a multi-level pattern (Tidwell, 2011, pp. 80-81). Having a multi-level navigation pattern is not always a problem, and can be an excellent feature if the users can circumvent some of the main pages, and go directly to sub-pages via navigation features such as sitemaps (Kalbach, 2007, pp. 63-65) that represents the entire website's navigation structure, or another menu-system that allows for subpages to be linked directly in the

global navigation, for example drop-down/fat menus (Tidwell, 2011, pp. 106-110; Kalbach, 2007, p. 75).

During the Think-Aloud test, one of the novices [N5] mentioned, when commenting on how hard he found the main website's structure to navigate, that he liked the menu in the footer of the site better than the top menus. He also mentions that he would have liked a drop-down menu, making it easier to get an overview of all the different subpages and their content.

Figure 28 – The footer menu of Feriecenter Slettestrand's main website. Note how the categories are almost similar to the labels in the top menu as seen in figure 27.

As seen in the figure above, the main website has a footer menu that works almost like a sitemap, except it does not have all the content and subpages in it like a complete sitemap normally would (Morville & Rosenfeld, 2006, p. 132). When comparing it to the top menu, it lacks the "Ferieboliger" and "Om os" labels, plus all the content from the top-most menu. This confused many of the participants, and they skipped using it because of it, and kept using the top menus instead. Expert participant [E2] even mentioned that he would have had removed the footer menu entirely (because he thought that the content of it was the same as the top menus, and thus only being redundant.) He did not discover that the menu content was not entirely the same.

One of the reasons we found that many of the participants skipped the footer menu can be explained with the layout of the main website. Most of the participants either found the footer menu very late in the process, or never found it at all. This is likely because they never had a reason to scroll down enough to discover it, as the other global navigation is always visible at the top, and the long content pages fill so much on the website, that the footer stays hidden most of the time. In the Heuristic Evaluation, we also noted that the menu in the footer is quite good, but can easily be overlooked because there is too much content (text, pictures, tables and local navigation menus) on all of the sites that the visitor often reads through, then goes back to the top menu that he is familiar with, instead of finding the menu at the bottom of the page.

The novice participants did however remark, more than the experts, that they found the amount of content, and the way of navigating it, to be too overwhelming for their liking. The experts tended to focus more on the positive aspects of the content, rather than thinking about ways to make it easier to navigate.

From a learnability perspective, there are some issues related to when participants chose to completely skip an essential part of the global navigation. The novice participants did especially bring up their concerns about the *too-many-choices* navigation, and how it confused them more than it helped them. However, as indicated by some of the participants and in the data from our Heuristic Evaluation, the footer menu could become a much better learnability feature, if it was easier discovered, and if the navigation options in it was consistent with the navigation options from the other global navigation so that they were comparable in options, allowing the users to learn the navigation features faster (Russell-Rose & Tate, 2013, pp. 84-86). It is not though, and we also noticed that two of the expert participants used the web browser's search function to search for keywords within the content of the sites of Feriecenter Slettestrand's main website, to solve the tasks quicker by completely circumventing the navigation features of the website. This could indicate that they were either extra focused on completing the tasks fast, could not be bothered to use the navigation features entirely, or a combination of both. None of the novices used this feature, and since we did not recruit participants based on their technical expertise (which in this case would be related to their ability to use web browsers efficiently) we regard it as a test variable and coincidence.

5.8 Evaluating Learnability through Clicks, Time and Task Fail-rates

As we measured how well the two participant groups did in terms of amount of time and number of mouse clicks spent during the Think-Aloud tasks (see figure 21 and 22 in the data presentation section), we discovered that the expert participants had a slightly lower task fail percentage than the novices, and used marginally less mouse clicks to solve the tasks. The novices also spent quite a lot more time solving the tasks, overall, than the experts did; the combined average time (for all tasks) spent for the novice participants is 725,8 seconds, and 586,8 seconds for the experts (23,7% faster).

The tasks that were used to measure this are tasks related to both Feriecenter Slettestrand's main website and mountain biking website, but the data is quite consistent throughout the entire test session for each participant group, making it harder to distinguish why exactly the domain experts did better overall. It can be argued that both groups learned the websites better and better for each task they completed, because when we compare the time spent and clicks used for each task, both groups use

much more time and clicks in average for the first few tasks, than the lasts tasks. This indicates that *some* learning takes place, possibly because of participants adapting to the navigation features and learning the different categories of content, which helps them solve the next task. It is hard to tell exactly how *much* learning takes place though, as the tests were only a single session, and for evaluating learnability, one of the issues is that it is hard to measure the results over longer time periods (Hornbæk, 2006, p. 93).

As a result, we see that participants with domain expertise (even though it is not relevant for all the tasks) solve tasks faster than participants with no domain expertise. They also spend less actions/mouse clicks and fail to complete tasks a little less (even though this number is marginal). The question is, if this is enough reason to always choose domain experts as participants, when evaluating learnability?

The Think-Aloud method is interesting as a test method for usability evaluation, as you cannot always be sure what kind of data you acquire when having completed the tests. You can hope, and you can create tasks that touch topics or design features you would like data and comments on, but you cannot be sure what the participants notice and which topics they verbalise their thoughts on. In our test data, the number of tasks solved by each participant group is nearly identical, and the variance between the two groups is spread over the entire spectrum of tasks.

The results only really vary significantly when comparing the time spent for each task, where novices spend more time on average. This is not necessarily a bad thing, when considering the test method; If you are measuring learnability, it could be argued that the more time the participants spend, the more time they have to verbalise their thoughts into tangible data, and they might also discover more content (and, as a consequence, more advantages or disadvantages of that content, which is useful data for evaluations). We see in our data, that a larger number of novice participants expressed their concern on the confusing global navigation menus than the experts did. They spent more time and clicks using those menus too, increasing the chance of bringing up excellent points for the data recordings and the evaluation.

The expert participants, we see in the data, had a tendency to use their own methods for finding the information they needed to solve the tasks – and it worked. They did complete the tasks faster on average, and used less clicks on average to do so. They also implemented new ways of solving the tasks, indicating what type of design features they would find useful for further development iterations of the site, or what type of usability features could help them utilise the websites even better

([E5] mentions that he was looking for a search function on the site, which does not exist at the moment). The experts did, as a consequence of solving the tasks faster, also spend less time on the test in total, reducing the amount of time they actually spent verbalising their thoughts, which *could* (especially when you have a low number of participants) result in poorer data, if the expert participants were completely focused on solving the tasks, rather than using the task as an incentive to think out loud (Nielsen, 1993, pp. 195-200).

By our experience, and by the data we collected, it seems like there are advantages and disadvantages to both types of participant groups. On one hand, the expert participants solved tasks more effectively in terms of speed and actions, but the novice participants were better to use the extra time they spent to add to the overall data collection. For evaluating the learnability attribute, it can be hard to argue which of these aspects is better, but having the participants spend extra time to verbalise their thoughts and state the reason for their actions, like we see the novices did better, is in many circumstances an advantage.

In the case of Feriecenter Slettestrand's two websites, and especially with the main website, several learning-hindering issues were discovered, particularly in the aspect of navigating the content of that site. The learning curve did go up, as the participants continued to solve tasks and started to learn the navigation features by experience, but implementing ways for the participants to always know "where they are" or not to get lost in the navigation features could improve the way user navigate the site. In the Heuristic Evaluation data (under point 9) we argue that a breadcrumb trail (Kalbach, 2007, pp. 60-63; Russell-Rose & Tate, 2013, pp. 197-199; Tidwell, 2011, pp. 121-123) showing the hierarchy of parent content pages could help users understand the navigation better, especially when there are so many subpages that can only be reached through main pages accessed from the top menu, or directly from the footer menu that often is not discovered. An example of a breadcrumb trail could look like:

Forside → Mountainbike → MTB Aktiviteter → Guidede ture

Having a breadcrumb trail, showing the parent and current page(s), could help the users find information or specific pages (again) faster than having to go to the top menu or front page each time the user is looking for something new or wants to go back to a previous visited page. The learnability curve would also be affected, giving the users a better chance to understand the structure of the entire website better, which is especially important as it is one of the problems that the test participants expressed their concerns about the most.

5.9 Determining User Satisfaction

As mentioned earlier, the criterion of *satisfaction* might be one of the hardest usability attributes to measure and evaluate. Learnability affects, to some extent, the users' perception of the system, and by that his satisfaction, but so do other attributes, such as effectiveness and efficiency. It would also be hard to deny, that satisfaction is a very subjective, and derived from the users' individual perceptions of what they find *satisfying* (Barnum, 2011, p. 12). Nevertheless, the satisfaction criterion can also, in a way, overrule attributes such as these, as satisfaction is directly related to the user's desire to use the system or product. If the user's desire to use it trumps eventual usability issues found in other attributes of the overall usability, you would be able to argue that the user is still somewhat satisfied, but that there are attributes that could still be improved, and heighten the user satisfaction and overall usability still.

Dimensions to *satisfaction* are many, but some of the most general guidelines include dimensions to what makes a product or system desirable to use, such as *effective, efficient, engaging, error tolerant* and *easy to learn* (Barnum, 2011, p. 12; Nielsen, 1993, pp. 33-37). In other words, are the users happy with how the system works, would they recommend it to a friend, and do they feel the need to use or have it? Since these questions are hard to measure objectively, many times the satisfaction (and eventual anxieties) can be studied by just asking the test participants about this (Nielsen, 1993, p. 209). We incorporated interview-style tasks in our Think-Aloud research design that inquired into these topics, which is also why we discussed the advantages and disadvantages of the concurrent and retrospective forms of conducting the method. We asked the participants, what two things they would change on each website and if the contents of the websites' lived up to their expectations of what websites like these should contain of information. Since we asked the participants in this way, and did not ask them to rate their experience or happiness with the websites, it is hard to summarise the data in a very practical way. We can, however, derive some of the answers and relate them to the dimensions that are related to the satisfaction criterion:

Determining Satisfaction: Novices

Most of the novices had issues finding the right content for many of the tasks they were asked to complete, and as we already discussed, they were generally slower and took more clicks to complete the tasks, compared to the expert participants. Being the slower participant group, it could indicate that the novices found the websites less manageable to navigate than the experts, which could inhibit their overall satisfaction of the websites.

All of the novices noted, that on the main website, there is a lot of content, text and text-quotes related to that content that they either "*ignore because it just looks like text with no purpose*" (cited from [N1] appendix 09, video 01 (17 minute mark)) or get lost in because they cannot get an overview of it and the information it contains. Novice participants [N1, N2, N4 and N5] also express their opinion on the front page of the main website, which they find to contain too much information that is useless to them.

The front page consists of, besides from the before-mentioned top and bottom main navigation menus, two large text quotes, a newsletter sign-up box, and then another type of menu that contain the labels "Aktiv i naturen", "Ophold", "Kursus" and "Restaurant". Being the first thing most participants put their attention to, when they visited the website the first time, almost none of them actually found it to be useful. The main labels cannot be clicked (even though many participants tried), and the four links underneath each label were in most cases for something that the participants were not looking for anyway, since they often link to *very* specific information, that only some target groups would relate to, or the labels are so ambiguous that the participants did not know what to expect would happen when they would click it, so they did not.

Figure 29 - Front Page menu of the main website

Accordingly, the novices did find a lot of hindrances for how well they were able to use the main website in in terms of effectiveness, efficiency, how engaging, error tolerant and easy to learn it was to them. They did, nevertheless, complete most of the tasks, and were able to give suggestions for how *they* would change many of the design features to accommodate their goals easier.

When asked about their expectations of the content, and if those expectations had been fulfilled, the common answer was that they did not find the actual content problematic, but it was rather a question of how that content had been organised and structured in ways that they found problematic. It was also mentioned how they would recommend Feriecenter Slettestrand to make the most out of their

physical environment, and try to introduce their website visitors to the beautiful nature and surroundings[24] through better visualisation of the nearby environment, for example by incorporating more evocative pictures or videos of the area, rather than the long texts that follow every subpage. Since the websites are both text-heavy, they wished that there were more ways that it appealed visually, using pictures or video to demonstrate the holiday centre's facilities, rather than long text paragraphs.

On the mountain biking website, four of the novices [N2, N3, N4 and N5] notice how there are more of this type of media, than on the main page, and the novices agree that the pictures invoke a mood that "*makes you feel like you know what you are going in for, and is more integrated*" (cited from [N2] appendix 09, video 08 (23 minute mark)) or that "*the website feels much more alive* [compared to the main website] *because of the pictures and interactive links*" (cited from [N5] appendix 09, video 18 (34 minute mark)).

Determining Satisfaction: Experts

Much like the novices, the experts also agreed that the mountain biking website, with its larger focus on the visual appearance (by including a lot more media, for example as background pictures) and more dynamic front page content (blog posts etc.) felt more "*action packed and dynamic*" (cited from [E4] appendix 09, video 11 (22 minute mark)) and that the "*idea of a blog and updated content on the front page is good*" (cited from [E3] appendix 09, video 7-2 (12 minute mark)).

They are also a bit more critical, when it comes to the visual appearance though: [E1] finds the mountain biking to be badly arranged because there are too many things *happening* on the entire website, which adds to his confusion, and makes it hard to get an actual overview of the content that is being presented. [E1, E2 and E5] all agree, that the pictures on the front page fills too much (it almost fills out the entire front page), and that important information is being pushed down at the bottom of the page because of it. The large pictures and content fields above it cause the issue, which is problematic because much of the information on the front page is quite critical and can be hard or impossible to find by just using the menu and exploring subpages.

The expert participants were also better to recommend which exact functions or information they needed, and where they would recommend it to be accessible. For example, the expert participants

[24] A few of the novices knew the area of the holiday centre.

used their domain expertise to specify what type of mountain biking information they would find useful to have on the front page: Experts [E1, E3 and E4] all mention how they would like relevant information, such as the current condition of the mountain biking track, information on bike rental, an updated list of upcoming events/races or a Facebook feed that links directly to the latest posts by Feriecenter Slettestrand's Facebook page[25] (as this was already his go-to choice for mountain bike updates from the holiday resort).

All of the expert participants were, because of their domain expertise, also able to more easily tell what type of relevant mountain biking information they expected to see on the two websites, and were more surprised if the content did not live up to their expectations. This include the before-mentioned updates on track conditions, but also how they would have liked a map of the track directly on Feriecenter Slettestrand's website (right now they only have a link to the map on The Danish Nature Agency's website[26]) and more specific information of the track (length, material, elevation etc.). Four of the novices also mention that they would like information on the track to be available, but are less specific.

Overall, the experts are more demanding when they were used to evaluate mountain biking specific parts of the websites. When asked to find a map of the track "Svinkløv Klitplantage", expert participant [E2] mentions that he would not even try to find this type of information on Feriecenter Slettestrand's website (even though he was already on the website), but would rather just use Feriecenter Slettestrand's *Strava* group[27], a social network for athletes, which already has this type of information included.

Participant [E2] also mentions that he would have liked a calendar function on the mountain biking website directly on the front page, because he would only really visit the site to find information on events and races that he could participate in. This is a tendency that we can also see in our Card Sorting results, where the expert participants would like to see relevant labels like "Nyheder", "Kalender" and "Aktiviteter" directly on the front page, much more than the novices would (see figure 7).

[25] http://www.facebook.com/pages/Feriecenter-Slettestrand/127013360707943
[26] http://www.naturstyrelsen.dk/publikationer/2009/dec/mountainbike-i-svinkloev-klitplantage
[27] http://app.strava.com/clubs/mountainbike-slettestrand-13260

5.10 Summary of Learnability and Satisfaction

A big part of making sure that there is an "*absence of frustration*" (Rubin & Chisnell, 2008, p. 4) in websites, as a metaphor for usability, is composed of various attributes. Learnability involves how well users can learn to utilise the website to their advantage and to fulfil their goals, and satisfaction is determined by how well the users perceive the website overall, based on their subjective criterions. One of the biggest issues concerning learnability on Feriecenter Slettestrand's websites, we found, is the way that content has been organised. For learnability to take place, the users need to *learn* how to use the websites to their advantage, but the learning aspect is hindered especially because of the *too-many-choices* navigation that the websites (especially the main website) is built upon. The main website contains, relatively, many subpages of content, and so it seems that the developers who built the website originally have tried to make sure that all of this content is accessible through many various menus and link under each subpage. During our usability evaluation, and by analysing the data it entailed, we found that participants, regardless of domain expertise, had a hard time making sense of the various menus and their labelling inconsistency, and so only really use one of the menu options regularly. The domain expert participants did however complete the tasks on average 23,7% faster than the novices, and were marginally more efficient on the number of actions they needed to complete those tasks (novices: 725,8 seconds, experts: 586,8 seconds). Even though they spent more time solving the tasks, the novice participants provided feedback that we found was just as valuable in usability evaluation contexts, as the feedback the experts provided. This finding contradicts some of the general guidelines of usability studies, which emphasize the importance of participants with expertise that fits the domain of the product being evaluated (Jenkins et al., 2003; Rubin & Chisnell, 2008, p. 115; Sova & Nielsen, 2003, pp. 29-31), but is supported by other studies that show the contrary, and had domain novice participants find as many (if not more) usability problems in an evaluation setting, and are therefore as useful, as expert participants (Karapanos et al., 2008; Kjeldskov et al., 2010).

When trying to determine the much more subjective attribute of user satisfaction, we found that the expert participants were more focused on exactly what type of information they expected to find, where to find it, and were more disappointed when they could not. This is especially the case, when they were used to evaluate parts of the websites that are directly linked to their domain expertise of mountain biking. This can be explained by how well their mental models (Barnum, 2011, p. 293; Benyon, 2010, pp. 32-34) were in compliance with the content, structure and functionality of Feriecenter Slettestrand's websites, where the domain experts definitely have more demands of what

exact type of information they want to match with their previous experiences of websites that relate to the same domain subject. On the other hand, the novices had a harder time trying to relate to many of the subjects of the websites, which consequently influenced the test data so that it was generally more restrained or unfocused when it came to the overall evaluation. It is important to keep in mind, that this result is based on a very subjective attribute of usability, and that the results could have been influenced by few, more eccentric participants who might have had very specific expectations of the websites and their content, compared to participants who might not have had any particular expectations, which might have skewed the test data towards the opinion and expectations of those few participants.

It is hard to determine *how* satisfied the participant groups were with the websites, but the novices might have the lowest level of satisfaction of the two groups, as indicated by their slower task solution times and more widespread data. The expert participants seemed to have a sharper focus in some areas of the evaluation, but they did not provide *as much* data. The problem will be to strike the right balance between the right amount of data you would want to collect during a usability evaluation, and how detailed you would like that data to be. Overall, both participant groups provided ideas that we agree with in our Heuristic Evaluation, and could be used to improve the websites' usability and Information Architecture, especially in terms of navigation, content and information seeking.

5.11 Errors and Safety

It is important to distinguish the difference between an error and a usability issue, as they are not the same thing. Errors are the outcome of a usability issues, rather than being the issue by itself (Tullis & Albert, 2013, p. 82). Imagine a user who wants to book a stay at a hotel, using the hotel's online booking system; He might chose the type of accommodation, how many days he would like to stay and which date he would like to check in, but because the booking system automatically showed the next month's dates, he might have picked a wrong date for his stay. This would inhibit his goal of booking a hotel room. Jakob Nielsen describes error situations as critical for usability because of two reasons: They represent situations where users are in so much trouble that they could become unable to use the system, and they represent situations where there is an opportunity to help the user understand the system better (Nielsen, 1993, p. 143). The less frequent these type of events occur for users, the better the overall feeling of being *safe* gets.

Since errors, and the overall safety of the system, occur somewhat fortuitously (and because we did not force them in our tests), we will not spend much time discussing this usability attribute. Another reason for not doing so, lies with how little difference there is when participants with varying domain expertise react to errors. Having the skills to be able to handle system errors (critical or not) effectively, is rather a technical expertise – and we did not recruit our test participants based on their technical expertise. We also did not encounter many errors during our evaluation tests, and the errors that we did encounter were often only experienced by a single participant by random choice. However, we will mention some of the more critical errors that we discovered during the Think-Aloud test and the Heuristic Evaluation, as they are still important for a usability evaluation.

The first of the two only errors we encountered consistently during the Think-Aloud tests has already been mentioned; the top menu of the main navigation is simply too small and hard to see, especially the top most menu (see figure 27). This is a design error, because that top part of the menu contains some of the labels that are most important for many tasks at Feriecenter Slettestrand's main website (*Kontakt*, *Booking*, *Priser*), but they were overlooked by many of the participants, regardless of domain expertise.

The other error we encountered consistently with our test participants is when there are links on Feriecenter Slettestrand's main website that links to external sites or documents (PDF files especially). Linking to a file or external site is not really an error or problem in itself, but on the main website, it is done in a way that makes it an error for many of the participants. The most critical

example many of the participants found (and we also found during the Heuristic Evaluation) is when an user wants to find the price for a specific holiday home, there is a link to a PDF file which would include the prices for the different holiday seasons (months/dates). Under the link to the PDF file, there is an image preview (Tidwell, 2011, pp. 263-266) of what the file contains, which looks like an ordinary calendar where the dates have been colour-indexed by price. There are several problems with this: Many of the participants tried to interact with the preview picture, because they thought it was the actual calendar to check for prices. When they realised it was not, and found out that it was just a preview, they were even more bewildered when they opened the PDF file, and the content of that file is not in consistence with how the preview picture made them expect it to look like. Instead, they are presented with a completely new type of price calendar, which looks nothing like the preview picture and what they were expecting to see.

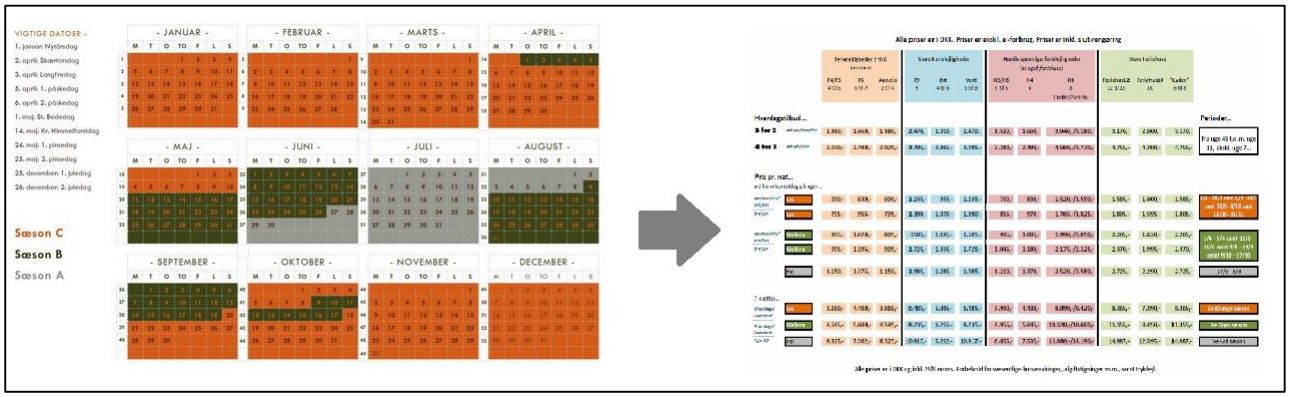

Figure 30 - The pricing calendars from the main website. To the left is the preview picture that users expect to see when they open the attached PDF file, to the right is what the file actually contains. Notice the big difference in colour choice and how the two types of calendars are structured.

Besides from confusing the users by showing an inconsistent preview picture of what the file contains, there is also no indication or warning that the user is about to open a PDF file. Some of the participants did not know that they were actually opening a completely new file (as it opened in the same browser window they were already using, instead of a new browser tab or window), and did not know how to go back to the previous site, until they realised that it was a PDF, and they had to use the browser's built in *go back* button, rather than the main navigation that obviously disappeared when they opened the file. If they had been warned about the fact they were opening a PDF file, this might not have been a problem at all.

We did find quite a lot more errors (critical and non-critical) during our Heuristic Evaluation, but since most of the test participants encountered only few, if any, of these errors during the Think-Aloud tests, we will not go through them here, as we cannot compare if and how the participants'

domain expertise affected their behaviour during their encounter with these errors. Those findings can be found in appendix 08 though.

5.12 Summary of Errors and Safety

Could domain expertise have had any influence on how participants reacted to the same types of errors, and could the difference have been compared? Maybe, but probably not. We do not have any data that would enable us to investigate this problem, but considering how the errors that we have mentioned above are mainly of a type that have to do with technical knowledge, rather than anything about mountain biking (or other domains), domain expertise probably would not play an important role. Technical expertise might have, if the results can be compared to those found in the study by Lazonder et al., where technical expertise was found to help participants locate specific information on websites faster, than those without technical expertise (Lazonder et al., 2000). If the same circumstances can be applied to error-solving, the results might be the same.

Nevertheless, the frequency of errors, and the feeling of safety on websites like Feriecenter Slettestrand's are still important attributes to the overall usability evaluation. In the two errors that our test participants consistently encountered, we have shown how small design changes would improve aspects of navigation and improve user expectation consistency.

If we were able to reiterate our Think-Aloud process, we would have incorporated tasks that deliberately would have tried to induce some of the critical errors we found in the Heuristic Evaluation, as it would enable us to compare how the two participant groups would react, and if those reactions would differ from one another depending on the level of domain expertise, even though most of the errors are preventable, or at least more understandable, by having a relevant technical expertise (Russel-Rose & Tate, 2013, pp 4-9).

5.13 Labelling

Creating the navigational labelling is an important process, as the visitors regard the labels as the structure of the website. Labels are a representation of a websites content and is used to make sure that users can find the information they are looking for quickly, without using time looking through the whole website (Morville & Rosenfeld, 2006, pp. 82-86). If the labels do not correlate with the mental models (Benyon, 2010, pp. 32-34) of the visitors, the intuitive navigation of the website will decrease (Kalbach, 2007, pp. 120-125). Labelling is one of the attributes that have a direct link to the intuitive design, which we presented in the theory section. How easily users navigate, explore and find the information they are looking for is related to the intuitive design of the website (Kalbach, 2007, pp. 34-37). Allocating time for developing navigational labels corresponding to the user mental models (Benyon, 2010, pp. 32-34) is therefore important and should be given a high priority. To understand our participant's mental models, we chose Card Sorting as our test method, as it gives insight into how the participants categorise content in their mind (Morville & Rosenfeld, 2006, pp. 106-108 and 255-259). To understand the mental models of the participants, we have used Card Sorting, as it gives insights into the mental models of its participants, for example how the participants sort and organise content in their heads (Spencer, 2009, pp. 10-13). The point of the method is therefore to choose participants who are (potential) users of a given website, as to understand their mental models and design labels and navigation that fits the mental models of the users (US Department of Health and Human Services, 2015a).

It is then interesting for us to analyse the data from the Card Sorting, as we have two very different participant groups, where one of the groups contain domain expert participants, who are potential users, and the other group consist of domain novices (who are the participants our hypothesis says can contribute, even though they are not considered the actual users or have the domain knowledge.) The literature then suggest that using participants who are not the potential users, will give results that could not be used for the design of the navigational labels, as they do not have the same mental models. To find out if this is the case, we will in the following section analyse how the domain experts sorted the labels in comparison to how the novices sorted the labels. This will show us whether the results from the two groups are similar or very different from each other, which can validate that in order to do Card Sorting, you will need to use potential users (domain experts).

To examine this, we will bring in selected data from the Card Sortings that we completed with the 20 participants (10 domain novices and 10 domain experts). We will analyse the data as to point out

similarities and differences between the two participant groups' sortings. Furthermore, we will discuss what changes those similarities and differences will bring to an evaluation and what the effect would be to the suggestions that an evaluation will contribute. In the first part of the label section, the sortings of the A-cards (the not mountain bike domain specific) will be analysed and in the second part the B-cards (the mountain bike domain specific) will be analysed.

5.14 A-Cards (Non-Domain Specific)

In this section, we will analyse the A-cards and see what differences there were between the experts and the novices. As these are common labels that are not referring to a specific domain (other than what is relevant on a website for vacations), we expected the sortings to be somewhat identical.

If we then look at the data from the Card Sorting, 40% of the novices want to be able to see a booking functionality directly on the front page (figure 3), compared to the experts, where only 20% want a booking functionality directly on the front page (figure 4). When looking at the category labels that the participants were asked to put as the headline for their sorted groups, the numbers start to resemble each other, as 30% of the novices and 20% of the experts wanted a category with the label "Booking" as headline. The experts sorted the label "Booking" beneath seven different headlines (seven different categories), while the novices agreed more and sorted "Booking" beneath five different labels. The focus for the novices to have "Booking" directly on the front page and the novices shared agreement on sorting "Booking" into fewer categories shows that they may have a higher focus on booking a vacation, than the experts do.

If we look further into the categories where "Booking" was the headline, the novices have 18 different labels located in the "Booking" category, while the experts only have 13 labels sorted into their "Booking" categories. If this is because the experts, who are interested in mountain biking, are not specifically interested in booking a vacation or other activities, but only want to go mountain biking and then return home, is difficult to say. The tracks around Feriecenter Slettestrand are not restricted to paying guests at the holiday center, but are open to all, making a one-day trip to the Slettestrand area a possibility for the mountain biking experts, as 8 of the 10 experts we used were living in Northern Jutland. The novices did on the contrary not necessarily see Feriecenter Slettestrand as a place for mountain biking, but rather a holiday destination, meaning their focus could then be on a stay lasting longer than a one-day trip to Feriecenter Slettestrand. The information need (Morville & Rosenfeld, 2006, pp. 33-35) of the two participation groups can therefore be seen as different from

each other as the novices focus on booking, where the experts focus less on this and more on what Feriecenter Slettestrand can offer as we see in the next part.

If we continue to look for differences in the data from the two participant groups, there was another more significant difference. The experts are more prone to sort cards into a category they labelled "Vi tilbyder", instead of using the label "Aktiviteter", which is the label the novices use as a headline (figure 31.) The cards that the two groups sort into the categories are in many cases the same cards; "Aktiviteter for børn", "Aktiv ferie for grupper", "Teambuilding", "Personaleevents" and "Indendørsaktiviteter" all have been sorted together with a high percentage by the groups, but under two different headlines. It can be difficult to separate the meaning of the two labels, but there is a big difference in how the two user groups used the labels. This can also be seen in the way the experts and novices sort cards into the category "Vi tilbyder" (figure 31 and 32), as the novices only have three labels with an agreement of 20% or higher. The experts on the contrary have sorted 27 labels into the "Vi tilbyder" category with an agreement of 20% or higher, whereas five of the labels have a 50% or higher agreement. It is interesting to see how big a difference there is between the two groups and how they use the label "Aktiviteter" and "Vi tilbyder". The labels have both been taken from Feriecenter Slettestrand's main website, as explained in the research design section (section 4.4), but it is only the label "Aktiviteter" that is located in the menu and can be interacted with. "Vi tilbyder" is a section header (Kalbach, 2007, p. 127) used to present various activities that Feriecenter Slettestrand are linking to in the menu "Aktiviteter".

Card no	Card name	VI TILBYDER
1	Gastronomikurser	10%
2	Tilbud på ophold	
3	Aktiviteter for børn	10%
4	Om os	
5	Attraktioner og oplevelser	10%
6	Kursuscenter	10%
7	Restaurant	
8	Aktiv ferie for grupper	10%
9	Vi tilbyder	10%
10	Råvarene bag maden	
11	Værd at vide	
12	Specialkost	
13	Booking	
14	Vin og gourmetaften	
15	Aktiviteter	20%
16	Familien Kronborg	
17	Find Vej	
18	Julefrokoster	
19	Aktiviteter i naturen	10%
20	Ferieboliger	10%
21	Indkøbsmuligheder	10%
22	Handicapvenligt ophold	10%
23	Åbningstider	
24	Teambuilding	20%
25	Boenheden Delfinen	10%
26	Selskaber	10%
27	Teambuilding aktiviteter	10%
28	Generelle Betingelser	
29	Ophold	10%
30	Menu	
31	Hjælpemidler	
32	Ugens aktiviteter	
33	På ferie uden din institution	10%
34	Personaleevents	10%
35	Kontakt	
36	Handicapvenlige lejligheder/huse	
37	Aktiviteter for alle	10%
38	Forplejningspriser	
39	Indendørsaktiviteter	10%
40	Teambuilding overnatning/forplejning	20%
41	Kursuscentrets faciliteter	
42	Forespørgsel	
43	Priser	
	Cards in this category	**21**

Card no	Card name	VI TILBYDER
1	Gastronomikurser	20%
2	Tilbud på ophold	20%
3	Aktiviteter for børn	50%
4	Om os	
5	Attraktioner og oplevelser	40%
6	Kursuscenter	40%
7	Restaurant	
8	Aktiv ferie for grupper	40%
9	Vi tilbyder	60%
10	Råvarene bag maden	
11	Værd at vide	20%
12	Specialkost	
13	Booking	20%
14	Vin og gourmetaften	
15	Aktiviteter	40%
16	Familien Kronborg	
17	Find Vej	10%
18	Julefrokoster	10%
19	Aktiviteter i naturen	40%
20	Ferieboliger	40%
21	Indkøbsmuligheder	20%
22	Handicapvenligt ophold	33%
23	Åbningstider restaurant	
24	Teambuilding	50%
25	Boenheden Delfinen	20%
26	Selskaber	22%
27	Teambuilding aktiviteter	40%
28	Generelle Betingelser	
29	Ophold	20%
30	Menu	
31	Hjælpemidler	40%
32	Ugens aktiviteter	40%
33	På ferie uden din institution	30%
34	Personaleevents	50%
35	Kontakt	
36	Handicapvenlige lejligheder/huse	10%
37	Aktiviteter for alle	30%
38	Forplejningspriser	
39	Indendørsaktiviteter	50%
40	Teambuilding overnatning/forplejning	40%
41	Kursuscentrets faciliteter	30%
42	Forespørgsel	10%
43	Priser	10%
	Cards in this category	**32**

Figure 31 - This is an extract from the Correlation tab in appendix 04, where we can see the comparison between the novices and experts when sorting cards into a category with the headline "Vi tilbyder". It is very noticeable that the experts (on the right) were much more prone to use "Vi tilbyder" than the novices.

Card no	Card name	AKTIVITETER
1	Gastronomikurser	20%
2	Tilbud på ophold	
3	Aktiviteter for børn	80%
4	Om os	
5	Attraktioner og oplevelser	50%
6	Kursuscenter	
7	Restaurant	
8	Aktiv ferie for grupper	50%
9	Vi tilbyder	20%
10	Råvarene bag maden	
11	Værd at vide	
12	Specialkost	
13	Booking	
14	Vin og gourmetaften	10%
15	Aktiviteter	70%
16	Familien Kronborg	
17	Find Vej	
18	Julefrokoster	20%
19	Aktiviteter i naturen	80%
20	Ferieboliger	
21	Indkøbsmuligheder	
22	Handicapvenligt ophold	
23	Åbningstider	
24	Teambuilding	50%
25	Boenheden Delfinen	
26	Selskaber	
27	Teambuilding aktiviteter	60%
28	Generelle Betingelser	
29	Ophold	
30	Menu	
31	Hjælpemidler	
32	Ugens aktiviteter	80%
33	På ferie uden din institution	10%
34	Personaleevents	50%
35	Kontakt	
36	Handicapvenlige lejligheder/huse	
37	Aktiviteter for alle	60%
38	Forplejningspriser	
39	Indendørsaktiviteter	80%
40	Teambuilding overnatning/forplejning	50%
41	Kursuscentrets faciliteter	10%
42	Forespørgsel	
43	Priser	
	Cards in this category	**18**

Card no	Card name	AKTIVITETER
1	Gastronomikurser	
2	Tilbud på ophold	
3	Aktiviteter for børn	30%
4	Om os	
5	Attraktioner og oplevelser	30%
6	Kursuscenter	10%
7	Restaurant	
8	Aktiv ferie for grupper	
9	Vi tilbyder	
10	Råvarene bag maden	
11	Værd at vide	
12	Specialkost	
13	Booking	
14	Vin og gourmetaften	
15	Aktiviteter	40%
16	Familien Kronborg	
17	Find Vej	
18	Julefrokoster	
19	Aktiviteter i naturen	40%
20	Ferieboliger	
21	Indkøbsmuligheder	
22	Handicapvenligt ophold	
23	Åbningstider restaurant	
24	Teambuilding	20%
25	Boenheden Delfinen	
26	Selskaber	
27	Teambuilding aktiviteter	20%
28	Generelle Betingelser	
29	Ophold	
30	Menu	
31	Hjælpemidler	
32	Ugens aktiviteter	40%
33	På ferie uden din institution	
34	Personaleevents	10%
35	Kontakt	
36	Handicapvenlige lejligheder/huse	
37	Aktiviteter for alle	30%
38	Forplejningspriser	
39	Indendørsaktiviteter	30%
40	Teambuilding overnatning/forplejning	20%
41	Kursuscentrets faciliteter	10%
42	Forespørgsel	
43	Priser	
	Cards in this category	**13**

Figure 32 - This is an extract from the Correlation tab in appendix 04. Here a comparison between the percentages of cards used by the novices (on the left) and the experts on the right) is seen in regard to the label "Aktiviteter". The novices are much more prone to use "Aktiviteter" as a headline than the experts.

Differences

The results from our Card Sorting show that if Feriecenter Slettestrand should point their website towards users with no mountain biking knowledge or interest, they should keep the label "Aktiviteter", as this is highly used by the novices. If Feriecenter Slettestrand would like to target their website at the mountain biker segment, they should then change the navigational label "Aktiviteter" to "Vi tilbyder", as the experts are more likely to use this as headline for Feriecenter Slettestrand's activities in the Card Sorting. This shows that there is a difference between the two groups in regards to a label that does not specifically has anything to do with the mountain biking domain. This is interesting, as we would expect both user groups to sort labels not referring to mountain biking in sortings that would resemble each other. They do however show differences in their way of sorting here. This is interesting, as we would expect the experts to be the "active" participant group wanting to know what activities they can participate in. The experts may however see the label "Aktiviteter", as referring to activities that they should participate in (that may or may not refer to mountain biking) and as they are already active mountain bikers, they may not be interested in other activities. They may only want to know, what facilities that Feriecenter Slettestrand is offering them in regards to mountain biking.

Similarities

There are also similarities between the two participant groups, as for example the label "Restaurant" is highly used by both groups as a headline in their sortings. This is however not that surprising, as this is a label that differentiates itself from the rest of the labels, as it is focused on food and what Feriecenter Slettestrand is offering in that regard. However, the experts do agree more, and have used six labels in the "Restaurant" category, with an agreement of 90% or higher, whereas the novices do not have any cards with that high agreement (they have four cards with 80% agreement as their highest). This shows that the experts have a more combined way of sorting their cards and agree more in regards to the "Restaurant" label. Both groups also use "Om os" as a headline for a category and both groups have a high agreement percentage. Again, we can see how the expert participants are more consistent, and have a higher level of agreement that the novices, as the experts have six cards with an agreement of 50% or more, compared to the novices who only have two.

As seen, both groups can agree on some of the same categories, where there is a high percentage of agreement, but both participant groups also show similarities when it comes to categories/labels with a low percentage of agreement. The label "Kursuscenter" is the cause of much confusion and low percentage agreement, as both groups do not have an agreement higher than 40% when "Kursuscenter" is used as headline. The novices have 20 different labels and the experts have 25 different labels located in the "Kursuscenter" category (figure 33). In figure 33, we can also see that three experts and four novices use "Kursuscenter" as a headline and ideally the cards sorted into these categories should therefore show some form of agreement, but they do not. Four novices use "Kursuscenter" as a headline for a category, but only six of the twenty cards in those categories are used by more than one participant. The three experts, who used "Kursuscenter" as a headline, only used five out of the nineteen labels more than once. This shows that none of the groups agree on which labels should be located in a category named "Kursuscenter".

When looking at the categories in which the label "Kursuscenter" has been sorted into, the experts have sorted it into five and the novices have sorted it into six. This again shows that both groups do not agree on where this label should be located, which indicates that the label is too weak and should be changed. The participants are not sure what the label entitles, as they sort it into categories like "Om os", "Aktiviteter", "Vi tilbyder", "Praktisk information" and "Personaleevents", showing a wide variety of sortings, based on the test participants' different ways of perceiving and understanding the meaning of that label. What can be concluded is that the label "Kursuscenter" is not clear enough in

the mental models of both participation groups, as they are sorting it in many different ways. This is in large perspective to the category with the label “Restaurant” as headline, which in the experts’ sorting has four cards with an agreement of 100%.

Card no	Card name	KURSUSCENTER
1	Gastronomikurser	10%
2	Tilbud på ophold	10%
3	Aktiviteter for børn	10%
4	Om os	
5	Attraktioner og oplevelser	10%
6	Kursuscenter	40%
7	Restaurant	
8	Aktiv ferie for grupper	10%
9	Vi tilbyder	
10	Råvarene bag maden	
11	Værd at vide	
12	Specialkost	
13	Booking	
14	Vin og gourmetaften	
15	Aktiviteter	
16	Familien Kronborg	
17	Find Vej	
18	Julefrokoster	
19	Aktiviteter i naturen	
20	Ferieboliger	20%
21	Indkøbsmuligheder	
22	Handicapvenligt ophold	20%
23	Åbningstider	
24	Teambuilding	10%
25	Boenheden Delfinen	10%
26	Selskaber	
27	Teambuilding aktiviteter	10%
28	Generelle Betingelser	
29	Ophold	10%
30	Menu	
31	Hjælpemidler	10%
32	Ugens aktiviteter	
33	På ferie uden din institution	30%
34	Personaleevents	10%
35	Kontakt	
36	Handicapvenlige lejligheder/huse	20%
37	Aktiviteter for alle	10%
38	Forplejningspriser	10%
39	Indendørsaktiviteter	
40	Teambuilding overnatning/forplejning	10%
41	Kursuscentrets faciliteter	30%
42	Forespørgsel	
43	Priser	
	Cards in this category	**20**

Card no	Card name	KURSUSCENTER
1	Gastronomikurser	10%
2	Tilbud på ophold	10%
3	Aktiviteter for børn	
4	Om os	10%
5	Attraktioner og oplevelser	
6	Kursuscenter	30%
7	Restaurant	
8	Aktiv ferie for grupper	
9	Vi tilbyder	
10	Råvarene bag maden	
11	Værd at vide	10%
12	Specialkost	
13	Booking	10%
14	Vin og gourmetaften	
15	Aktiviteter	10%
16	Familien Kronborg	10%
17	Find Vej	10%
18	Julefrokoster	
19	Aktiviteter i naturen	
20	Ferieboliger	10%
21	Indkøbsmuligheder	10%
22	Handicapvenligt ophold	
23	Åbningstider restaurant	10%
24	Teambuilding	10%
25	Boenheden Delfinen	10%
26	Selskaber	
27	Teambuilding aktiviteter	10%
28	Generelle Betingelser	10%
29	Ophold	20%
30	Menu	
31	Hjælpemidler	10%
32	Ugens aktiviteter	
33	På ferie uden din institution	10%
34	Personaleevents	10%
35	Kontakt	
36	Handicapvenlige lejligheder/huse	10%
37	Aktiviteter for alle	
38	Forplejningspriser	
39	Indendørsaktiviteter	
40	Teambuilding overnatning/forplejning	10%
41	Kursuscentrets faciliteter	20%
42	Forespørgsel	10%
43	Priser	10%
	Cards in this category	**25**

Figure 33 - The novices’ (left) and experts’ (right) sorting with the headline "Kursuscenter".

This tendency could indicate that the label “Kursuscenter” is both confusing and does not precisely tell neither of the participation groups what that label entitles. The label is too vague for both participant groups and should be changed in order to be useful for the two groups (Morville & Rosenfeld, 2006, pp. 83-86). In an evaluation with either one of the groups would therefore lead to a change. To overcome this you can use focused labels (Kalbach, 2007, p. 128), where the aim is to narrow down the label, without specifying it to a degree where it does not include all of the under lying content.

Our recommendation would then be to rename the label “For virksomheder”, as the content is referring to activities for companies (teambuilding, rooms for courses, overnight stays etc.) The content is not aimed at holiday visitors or mountain bikers, but instead focused at companies and other institutions that could have an interest in using Feriecenter Slettestrand’s facilities. To ensure that the visitors who are not representing a company can easily avoid this labels’ content.

What the Card Sorting is not telling us is that some of the participants thought that the label "Kursuscenter" indicated Feriecenter Slettestrand's main building. This we found out doing the Think-Aloud, as participant would look for the opening hours of the reception. "For virksomheder" is therefore in our opinion a better label telling visitors that the content is only relevant, if they are representing a company.

5.15 Summary of A-cards (Non-Domain Specific)

In the above the cards from card sorting A has been analysed, which were the cards with labels from Feriecenter Slettestrand's main website not including the mountain bike specific labels. We made this Card Sorting as to see how the two groups sorted these common labels, with no direct link to mountain biking. We were expecting these sortings to be somewhat alike, as the mountain biking domain was excluded from this Card Sorting. We did however find that there was a significant difference in a few of the categories made by the two groups. The interest in making sure that a booking functionality was present on the front page was in higher regard for the novices than the experts. The experts' focus was not on booking but instead on what Feriecenter Slettestrand could offer them in regards to the "Vi tilbyder" label. The novices did almost not use this label as a headline and those who used it, did not agree on which labels that should be sorted into a category with the label "Vi tilbyder". The label "Aktiviteter" was much more popular among the novices and the labels that the experts sorted into the category "Vi tilbyder" were the labels the novices used in their "Aktiviteter" category.

This was very interesting, as it would have given two variations of the main website; one where the label "Vi tilbyder" was implemented and one where the label "Aktiviteter" was implemented (or kept, as this was already the navigational label for this category on the main website.) Labels should always relate to the users and their terminology (Kalbach, 2007, pp. 123-127) and using the mountain biking expert group in a usability evaluation (without the novices) would therefore have led to choosing "Vi tilbyder". However, there were also similarities as both groups disagreed on the sorting of the label "Kursuscenter". Both groups used the label "Kursuscenter" as a headline for a category and as a subjacent label in various categories with different headlines. The meaning of this label was therefore too ambiguous (Kalbach, 2007, p. 123), as the two groups did not sort it in an agreeable way. The suggestion would then be to rename the label when using either of the two groups, showing that a label can be ambiguous in both groups.

What the analysis of how the participants sorted the A-cards has shown us that even though the labels may seem common, there can still be differences between groups that have expertise on a given

domain and know that they have been chosen as participants for a test because of this knowledge and a group that have been assembled with participants with no form of common interest or specific knowledge on a given domain.

The data from the Card Sorting with the A-cards lead to two navigational labelling systems (one for the experts and one for the novices). The results from the expert's sortings, where there are three or more experts who have sorted the same label into a category, lead to a navigational labelling system with the following labels:

Aktiviteter	Om os	Kursuscenter	Restaurant	Vi tilbyder	Handicapvenligt ophold

The label put on the front page by the experts, where four or more experts agreed, was "Ugens aktiviteter" (taken from figure 4, where four or more experts have agreed on a label). The results from the novice's sortings, where there are three or more novices who have sorted the same label into a category, lead to a navigational labelling system with the following labels:

Aktiviteter	Om os	Kursuscenter	Restaurant	Booking	Værd at vide

The labels put on the front page by the novices, where four or more novices agreed, were "Aktiviteter", "Booking" and "Handicapvenligt ophold" (taken from figure 3, where four or more novices have agreed on a label). This shows two variations of the navigational labelling system, where especially the labels put onto the front page is different between the two participant groups, as the experts only agreed on one (four or more experts agreeing on the same label). The two systems do however also show labelling systems that resemblances each other, as four of the labels used by both participant groups are the same in the navigational labelling system. There would however be redundancy in the experts' labels, as they use both "Aktiviteter" and "Vi tilbyder". A decision should then be made to remove "Aktiviteter", as it has the least agreement in the expert group.

In the next part of this section we will look at the B-cards from the Card Sorting, which are the cards relating to mountain biking and have been given labels from Feriecenter Slettestrand's mountain biking website and mountain biking content from the main website.

5.16 B-Cards (Domain Specific)

In this section we expected, before doing the Card Sorting test, that the results from the two participant groups would be noticeable different, as the cards were now referring to mountain biking. This is also what happened with some of the labels, but there were also similarities that make both participant groups equally useful in an evaluation. The label "På Sporet" was used two times by the experts and two times by the novices for a headline. The label "Ruterne" was used four times by the experts and two times by the novices for a headline. This shows that neither of the participants in the two groups agreed on either of the labels. The novices also used the label "Stier og spor i området" in regards to the tracks. The three labels refer to the same content, but are used inconsistently throughout the websites. This is also one of the critiques in the Heuristic Evaluation (appendix 08), as there is no common terminology used on Feriecenter Slettestrand's websites that is used to describe their mountain biking track. Instead, they use a variety of terms, such as "På Sporet", "Ruterne" and "Stier og spor i området", which makes it difficult for the visitors to know exactly what to look for. Feriecenter Slettestrand needs to be more consistent when referring to their track.

The sortings made by both groups do not give a decisive answer to what label should be used, but the novices have used "På sporet" more than "Ruterne", and the experts use the label "Ruterne" more than "På sporet". In an evaluation the suggestions would therefore be different dependent on which participant groups is being used, as the novices lean towards "På sporet" and the experts lean towards "Ruterne". The differences in agreement are however minimal. The interesting thing is that both groups see it as relevant to have a category for the track and that it should contain content like the condition of the tracks, an interactive map and video. This shows that even though the novices do not have any domain expertise on mountain biking, they are still able to presume that this is an important part of mountain biking. A suggestion made individually from the two groups would therefore be to have a navigational label called one of the three labels describing the tracks.

Card no	Card name	MOUNTAINBIKE
1	Teknik og taktik	40%
2	Bloggen	
3	Undervisning	30%
4	På sporet	40%
5	Nyheder	
6	Løb og events	20%
7	Køb gavekort	10%
8	Nyhedsservice	10%
9	Mountainbike	80%
10	Kalender 2015	
11	Aktivitet på træningsbanen	20%
12	Ruterne	40%
13	Find vej her	
14	Stier og spor i området	40%
15	Telefon og email	
16	Se video fra sporet	40%
17	Derfor kører vi MTB!	50%
18	Slettestrand.dk	
19	Sporets tilstand	30%
20	Kontakt	
21	Interaktivt kort	20%
22	Udlejning	50%
23	Aktiviteter	
24	Team MTB Slettestrand	30%
25	Ophold	20%
26	Tekniktræning	40%
27	Baggrund for MTB Slettestrand	30%
28	Kursus og events	20%
29	MTB Camp	30%
30	MTB Kurser	30%
31	MTB Udlejning	50%
	Cards in this category	**23**

Card no	Card name	MOUNTAINBIKE
1	Teknik og taktik	40%
2	Bloggen	
3	Undervisning	20%
4	På sporet	20%
5	Nyheder	
6	Løb og events	20%
7	Køb gavekort	
8	Nyhedsservice	
9	Mountainbike	40%
10	Kalender 2015	
11	Aktivitet på træningsbanen	20%
12	Ruterne	20%
13	Find vej her	
14	Stier og spor i området	20%
15	Telefon og email	
16	Se video fra sporet	20%
17	Derfor kører vi MTB!	40%
18	Slettestrand.dk	
19	Sporets tilstand	20%
20	Kontakt	
21	Interaktivt kort	10%
22	Udlejning	30%
23	Aktiviteter	10%
24	Team MTB Slettestrand	40%
25	Ophold	
26	Tekniktræning	40%
27	Baggrund for MTB Slettestrand	40%
28	Kursus og events	
29	MTB Camp	40%
30	MTB Kurser	40%
31	MTB Udlejning	40%
	Cards in this category	**20**

Figure 34 - Here the cards sorted into the "Mountainbike" category by the experts is presented to the left, and novices to the right.

However, there were also noticeable differences between the two groups and as seen in appendix 04 (*Correlation* tab in spreadsheets concerning B-cards) the experts agreed on fewer categories than the novices, as they only made 13 categories, compared to the novices who made 18 categories. One of the reasons for the less agreement in categories for the novices can be seen in the way they used the "Mountainbike" label.

The expert participants showed a great interest in using the "Mountainbike" label as a headline, whereas the novices are not as interested. As seen in figure 34, the experts used it eight times and novices used it four times. 80% of the experts used "Mountainbike" as a headline making that the most used by both groups. This could be an indication that the label should be used when using the mountain biking experts as participants in a usability evaluation, but when we look at the number of cards used in that category and small percentage in agreement, it does however show "Mountainbike" as a catch-all label (Kalbach, 2007, p. 127), where the participants put everything they have a difficulty in sorting into, as they have nowhere else to sort it to. The category ends up being a more general category, where the participants put those labels that they cannot relate to other labels into. This makes it difficult to give a useful suggestion as to how the label should be used. The sortings have labels like "Ophold" and "Gavekort", and "Udlejning" and "Teknik og taktik", which show that the label itself is too ambiguous for both groups. We did however also take the "Mountainbike" label

from the main website, where it is used as the main navigational label for the mountain bike content, where it makes sense to use it. If it used for a website only containing mountain specific content like Feriecenter Slettestrand's mountain biking website, this label is however not useful, as all of the content on that website can be sorted into a mountain biking category. The label is too broad to be used as a navigational label.

5.17 Summary of B-cards (Domain Specific)

In the above the B-cards have been analysed, as to see similarities and differences in the way our participant groups sort labels regarding the mountain biking domain. This was done in order to see how important it is to use participants with domain on a given subject, when doing a Card Sorting. We were expected this part of the Card Sorting, in comparison to the A-cards, to show large differences in the way the two groups sorted the cards, but we were also surprised to see that in some parts of the Card Sorting, there were an agreement between the two groups. Both groups made sortings that showed the relevance of having a navigational label leading to content in regard to the tracks and information on those. In the Think-Aloud tests, the novices did not show interest in the tracks and their condition, but here in the Card Sorting some of them did however still make a category for the purpose of giving information regarding the tracks.

The experts did however agree more on fewer categories than the novices, showing more unison in the domain expert group. The categories made by the novices were more spread out, which give less useful data for suggestions towards a better labelling system. The results from the experts' sortings, where there are three or more experts who have sorted the same label into a category, lead to a navigational labelling system with the labels:

Aktiviteter	Mountainbike	Kontakt	Praktisk information	Ruterne	Team MTB Slettestrand

The labels the experts chose for the front page were "Nyheder", "Kalender 2015", and "Sporets tilstand" (taken from figure 5, where four or more experts have agreed on a label). The results from the novice's sortings, where there are three or more novices who have sorted the same label into a category, lead to a navigational labelling system with the labels:

Aktiviteter	Mountainbike	Kontakt	Praktisk information	Kalender 2015	Nyheder

The labels the novices chose for the front page were “Kursus og events”, “Kontakt” and “Ophold” (taken from figure 6, where four or more novices have agreed on a label).

This shows two variations of the navigational labelling system, where especially the labels placed onto the front page are different between the two groups. It does however also show labelling systems that resemblance each other, as four of the labels used for categories are the same. The percentage of agreement in the two groups may vary, as the novices’ data were more spread out, but looking at the categories that stand out, the same four cards were chosen by the two groups. This shows that even though using participants that are not domain experts, a result can still be presented that resemblances the results by a group with expert domain knowledge, which is in contradiction to Rubin and Chisnell (2008), who claim that results can only be valid if the group of participants are representative of the intended user group and their abilities (Rubin & Chisnell, 2008, p. 115). The labels put on the front page by the two groups are however not the same, which corresponds with Rubin and Chisnell (2008) and for that given task (of choosing labels for the front page) the results from the domain novices do not resemblance the domain experts.

5.18 Navigation

In this section, we will look at Feriecenter Slettestrand's websites from a navigational perspective, both the advantages and disadvantages of its existing navigation, but more importantly how test data from the completed evaluation tests could help us to understand the navigation needs of the two participant groups, and if their domain expertise could have influenced these results.

Navigation, even though it is not normally directly connected to the usability term, but rather an important factor in a system's Information Architecture, plays an important role when trying to limit the level of frustration and confusion (Morville & Rosenfeld, 2006, pp. 115-116) – something that one could argue is one of the of the main goals of a usability study. To investigate how the navigation attribute is being affected by domain expertise, we will use data from both the Card Sorting and Think-Aloud tests.

5.19 Navigation as a Part of Intuitive Design

Navigation attributes to the overall usability of system or website by being a major factor of the *intuitive design* attribute of usability which seeks to deploy "*a nearly effortless understanding of the architecture and navigation of the site*" (U.S. Department of Health and Human Services, 2015b). Such an understanding requires a navigational system that the users are able to quickly learn and understand (something we already discussed in the section on Learnability and Satisfaction, section 5.6), so that they are able to use it efficiently to complete their goals on the websites. If the navigation becomes a problem (by being hard to use or understand) it just becomes an impediment for progress, and the user can get lost – something that is associated with anger, fear, frustration and confusion (Morville & Rosenfeld, 2006, p. 115).

Website designer and user experience specialist, Jenifer Tidwell, describes navigation as a metaphor for getting around by commuting – something that is necessary to get to where you want to, but also something that takes up resources (time) and is dull. She thinks that "*The best kind of commuting is none at all*" (Tidwell, 2011, p. 77), meaning that the more convenient and "within reach" the navigation is, the better the experience users *can* have. This also means, that if navigation is a factor of an intuitively designed system or website, it should also follow principles of minimalist design to reduce the amount of redundant information that is irrelevant, or rarely needed (Barnum, 2011, p. 62; Nielsen, 1993, pp. 129-132), as this can help users understand the hierarchy and design of the navigation systems more easily, than if they had to recognize many different labels or so much

information that it becomes hard to distinguish one element from another. Of course, things can get *too* minimalistic, and can become just as problematic if the available navigation choices are too limited or restricted.

5.20 Navigation Systems on Feriecenter Slettestrand's Websites

As we demonstrated in the section on Learnability and Satisfaction (section 5.6), the global navigation menus of Feriecenter Slettestrand's main website is very inconsistent, and includes too many choices for most participants to get an overview of how they can utilise the navigation systems most effectively. This was in effect with both participant groups, independently on their domain expertise. What we saw, was that most of the participants chose to only use the top menu of the global navigation, and generally avoided using the menu in the footer, as it was inconsistent with the choices available in the top menu. Both participant groups also used the local navigation menus (Morville & Rosenfeld, 2006, pp. 124-126) as a means to navigate the information found under the immediate subpages.

To circumvent the issues with the existing footer menu, which almost was not used by any participants, but still make it easy to get an overview of almost all the content on the websites, we suggested a drop-down/fat menu (Kalbach, 2007, p. 75; Tidwell, 2011, pp. 106-110) to replace the top and footer menus and could also, to some degree, reduce the options of choices in local navigation menus, by organising the websites' content into a single menu with a few, but meaningful and distinguishable choices, and with the relevant content and subpages sorted under each main choice.

Figure 35 - Example of a drop-down menu used on Louis Nielsen's website. The example demonstrates how this type of menu allows many subpages to be sorted into meaningful categories. Picture from http://www.louisnielsen.dk.

In the section concerning labelling (section 5.13) we examined which labels would be most efficient in terms of meaning and use contexts, but for this section we will look at which content categories would be most sensible to use in an eventual redesign of the navigation systems of Feriecenter Slettestrand's websites.

5.21 Choosing Content Categories for the Main Navigation

In the Card Sorting tests, when the participants were asked to pick the card labels they found important enough to use on the front page (for example as the menu choices in a drop-down menu), we see how the novice participants agree more on which A-cards (non-domain specific cards) to use than the expert participants (figure 3 and 4), as they have more single cards with a higher percentage of agreement than the experts. The opposite is seen, when comparing how the participants chose which B-cards (mountain bike domain specific) to use for the front page, where the experts have a higher agreement, than the novices (figure 6 and 7). This can be seen by how the expert participants have a lower number of unique cards chosen for this task. The novices do have one card which was chosen by five participants, which is more than any card for the novices, but the novices have a higher count of total labels too, indicating more disagreement between the novices, than the experts.

Consequently, we can see that there is a clear difference in how domain expertise seems to influence the way that participants value the importance of content; the novices seem to agree more, when it comes to non-domain specific content, which is content that most participants probably have very individual mental models of *how* to value the importance of (Russel-Rose & Tate, 2013, pp. 24-25). When compared to the domain-specific B-cards, the expert participants, who have a more nuanced and experienced mental models of that domain subject, show a more consistent and focused choice in card importance. This can be seen by how they only picked nine B-cards for the front page (cards picked by only one participants are omitted, both for novices and experts), compared to the 13 cards chosen by the novices.

In terms of which content the participants found redundant or unnecessary for the websites in general, there was not much difference in how the two participant groups valued the cards. The expert participants deemed more A-cards unnecessary or redundant, compared to the novices (see figure 9 and 10), and with the B-cards, the experts were much more specific than the novices, as they agree much more on which cards to deem unnecessary/redundant. Only two of the novices agreed on the same B-card, and the rest of the choices were unique to one novice participant (see figure 12).

When comparing the total sum of cards chosen for the front page (by domain experts and novices combined) for the A-cards (see figure 5), we can see that overall, the level of card value agreement is a bit lower compared to the total sum of B-cards chosen for the front page (see figure 8). For the non-domain specific A-cards, the two most popular cards have only been chosen by six participants (out of all 20), but for the domain specific B-cards, the two most popular cards have been chosen by seven participants, and the third and fourth most popular cards have been chosen by six. However, for the A-cards, there is a higher level of agreement on *which* card should be used for a front page: A total of 17 unique A-cards were chosen, compared to 22 unique B-cards. This might not seem like a big difference, but when taking into consideration that there are a total of 43 A-cards, and 31 B-cards, it means that for the non-domain specific A-cards, only 39,5% of them were chosen by the participants for the front page task, whereas 71% of the B-cards were chosen for the same test, indicating that there was more disagreement overall for the B-cards, even though the expert participants were more focused in card choices for the B-cards, than they were for the A-cards.

To recapitulate, we see that for the non-domain specific card choices, the novice participants have a higher level of agreement than the expert participants – not by much, but enough to say that it is noticeable and could have an impact on a redesign, if the data were to be used for that. However, the expert participants show a more focused choice in cards and a higher level of agreement, when the task was using the domain-specific B-cards. In terms of which cards the participants found redundant or useless, not much difference in value is to be seen, except for when the expert participants sorted the B-cards, where their expertise is useful for being much more specific, and show a higher level of agreement than the novices did.

When the total number of cards chosen for the front page are added together, regardless of domain expertise, (see figure 5 and 8) we can also see that even though a lower percentage of unique cards were chosen for the test with non-domain specific A-cards, than the domain specific B-cards, the test with the B-cards has a slightly higher level of agreement. This is interesting because it tells us how domain-expertise *can* affect data results; the test with the A-cards was not domain-specific. Therefore, it should be expected that the results for the tests with A-cards would be somewhat alike between the two participant groups – which they were. But for the domain-specific tests using the B-cards, the novice participants have a considerably harder time agreeing on the value of cards, compared to the experts who, with their better mental models on that specific subject, have an easier time relating to and setting the domain-specific cards into a perspective that is useful to themselves and other expert users.

In the Card Sorting task where the participants were asked to choose a label for each of their sorting categories, the results are almost similar to the tasks described above, when focusing on how domain expertise affects the data from the test. In figure 15 and 16, we can see how there is not too much difference in the level of agreement within the two participant groups for the test using the non-domain specific A-cards. Each group has three or four cards that were popular within that group (5+ participants agreeing on the same unique card), and then three to four cards with medium agreement (3-4 participants agreeing on the same unique card). The novices have a higher number of unique cards with six or more participants having chosen those cards, but the experts are the only group with a unique card that all 10 of the expert participants used as a category label.

For the same test, but with the domain-specific B-cards (see graph 18 and 19) we can see the same pattern as the tasks before this one; the expert participants seem to have had a slightly higher level of agreement. They have two cards which over five participants used, and generally they seem to have chosen the same cards more times, compared to the novices. The novices' data is much more scattered, and they have used many more unique or participant created labels for their categories, indicating that they might have had issues using existing card labels. For the expert participants, only two participants have used a unique or participant created label (compared to 10 for the novice participants), showing that they have had less issues with using the existing content and labels as category headlines for their sortings. This could be because of better ability to divide the B-cards into categories that are much more nuanced (as those cards almost all are related to the mountain biking domain in some way), where the novice participants might have acted more on instinct, based on a lower or non-existing domain expertise, which has resulted in more scattered data results.

5.22 Navigation Design and Domain Expertise

We have already mentioned how a drop-down menu could help circumvent some of the issues that the participants experienced during the use of Feriecenter Slettestrand's websites, and how such a navigation is also supported by data from the tests. But choosing one type of navigation is not enough – it needs to be decided what other navigation mechanisms might be useful or could contribute to how users solve tasks or complete goals (Kalbach, 2007, p. 55). We also mentioned how a breadcrumb trail (Kalbach, 2007, pp. 60-62; Tidwell, 2011, pp. 77-79) could help users always know *where* they are, and how they can get back to content they have already visited.

What we have focused on in this section, so far, has rather been how the participant groups' domain expertise influenced the data we collected during the evaluation tests. Now we will enlarge on what

this means for the *design* of the navigation systems, and how the two different participant groups' data sets could have influenced the design of the navigation. This discussion will also be used as an introduction to another discussion concerning the use of participants with, or without, relevant domain expertise in usability evaluations overall, based on our findings from this and the other sections of our analysis.

Let us imagine that we had completed the same evaluation tests (primarily the Card Sorting and Think-Aloud) with the same tasks, but with only one of the participant groups, and only for the sake of designing a new navigation scheme for Feriecenter Slettestrand's websites. How would it influence the design choices, if we had only used domain novices, or only domain experts?

First off, we have not gone very much into detail on how we would redesign the existing navigation scheme if we were to do it now (as that is not the primary focus of our hypothesis), but as we have described, we can see how the domain expert participants seem to agree more on which content (represented by cards) they found to be most important, and how that content should be grouped in categories - but mainly when the focus of the test was revolving website content that had to do with the domain in which they were experts. For the content that was not bound to one single domain (primarily the tests that used A-cards), we found that the data does not vary that much depending on domain expertise. That is an important observation, as it can help us illuminate in which contexts it would be more useful to include participants that are both domain experts and novices, and where only one type of participants could yield the best data.

Our original assumption was that when evaluating websites where there is not only *one* target group, but several target groups in a single context (just like Feriecenter Slettestrand's), participants that do not fit the ideal recruitment description for those tests might still yield very relevant and useful data for usability evaluations. Had we only used novices for evaluating and redesigning the navigation system, the final design would probably be somewhat scattered, such as we see the novices' data is. That does not necessarily mean that the final navigation design would be bad, but it would have been influenced by data that is less precise in the mountain biking domain, but maybe have appealed to a larger, less defined target group of users. However, for the part of the navigation design that would not be tied to one specific domain, the results would not have changed that much – and since the context of this evaluation is constructed by several domains, the biggest issue might in reality be the fact that it is nearly impossible to find participants that would fit every target group's domain. If that is the case, our data shows that using participants with no particular domain expertise is just as useful,

if not even a little better, than using participants that fit the domain description of *one* of the domains that the overall context is built upon. Consequently, the redesign proposals for a new navigation system would inevitably be based on the data that the participants were used to accumulate, domain experts or not, and the proposed navigation system design would be affected in the sense that domain expert participants' data would allow for a more specified or subjectively appealing design, that would fit users with the same domain expertise, but not necessarily *other* or *generic* users.

This is because the attributes of usability as a term is very dependent on the context in which the system is being used. There is a big difference in how a person would look for vacation homes (prices, dates, housing types) for his whole family and for recreational purposes, than how a mountain biker would find information on a specific race and sign up for that race, especially since both tasks can be completed on the same website in Feriecenter Slettestrand's case. Both tasks are important, but they can be achieved in very different ways, and designing a navigation system that support both tasks, and still has a high level of usability for both persons, will be difficult (Morville & Rosenfeld, 2006, pp. 118-120).

Nevertheless, we have shown that even though test participants might lack a specific domain expertise, they can still be used to gather useful data for usability evaluation in some contexts and for target groups that those participants fit in already. Morville and Rosenfeld call this problem a *multidimensional puzzle* (Morville & Rosenfeld, 2006, p. 247), as these types of evaluation and redesign proposals often cannot be completed satisfactory by only looking at them from one single perspective, but should be considered from more than one point of view. Jakob Nielsen also discusses this problem, by showing the three main dimensions of what he calls the "user cube". The three dimensions include *users' varying experience with the system* that is being used, with *computers* (what we described earlier as technical expertise), and with *the task domain* (what we call domain expertise) (Nielsen, 1993, pp. 43-48). He suggests that it is best to find a safety margin of sorts (Morville and Rosenfeld call this *flexibility in navigation* (Morville & Rosenfeld, 2006, pp. 120-123)), which would allow all types of users to utilize the same system, but for different tasks, independently on their experience within each of the three dimensions.

5.23 Summary of Navigation

Navigation is versatile. The design of a navigation system for websites is very dependent on the context in which the users actually use it, which tasks they are trying to solve, what their goals are and how much experience within different domains and dimensions they have.

In this section concerning navigation, we have illustrated how test participants with domain expertise within the domain they are being used to test, can help test facilitators accumulate more precise data than with participants with no relevant domain expertise, but mainly in situations where the participants fit the domain context. In other contexts, for example when there are several domain contexts, domain novice participants seem to be able to deliver evaluation results that are on par with the results of domain experts, even though the test tasks are identical. These results are consistent with the results in Bednarik and Tukiainen's study, where expert participants were found to be able to process more information and deliver more precise data (Bednarik & Tukiainen, 2005), and also comparable with the results from the study by Dou et al., where domain expertise was found to be relevant and useful, but mainly in evaluation situations where the domain expertise fits the overriding domain context of the product being evaluated (Dou et al., 2009).

Since it seems like the *one solution* for designing navigation for contexts with several domain contexts does not exist, maybe the best possible action, when evaluating navigation as a part of the overall usability, is to include more than one type of participant or participant group, based on the existing domain contexts and target groups of the system, rather than only focusing on one type of participant. This would allow facilitators to collect data that is both relevant to exact domains, but also ensure that they are able to collect data that is not directly related to a specific context, but a rather general (or scattered) aspect of the system or website overall.

In the last section, we will discuss how test participant domain expertise could be used, or at least how we consider it should be reflected upon, in usability and Information Architecture evaluation processes, based on our findings in the previous analysis sections.

5.24 Discussion and Summary

In the previous parts of the analysis section, we have examined how the test results accumulated by our two participant groups resembled each other, and especially tried to focus on some of the important areas where they do not. We have done this in an effort to investigate how test participants' domain expertise influences the results of a usability evaluation process, where the context for the evaluation revolves around websites with several domain contexts, hence also having several distinguishable target groups. The reason for doing this, lies with our original curiosity of investigating how participant domain expertise is factor that is important to consider, when recruiting participants for evaluation studies, whether it is usability or Information Architecture that is being evaluated.

5.25 The Influence of Participant Domain Expertise in Evaluation Processes

Our hypothesis was formed from our own experiences with user testing in earlier work, where we have discovered that participants with no particular domain expertise still were very useful for evaluation processes, but also from the fact that our understanding did not entirely match what we learned was the prevalent opinion and practice, when conducting this type of evaluation processes. In the relevant literature (Bednarik & Tukiainen, 2005; Botella et al., 2014; Dou et al., 2009; Karapanos et al., 2008; Kinney et al., 2008; Kjeldskov et al., 2010; Lazonder et al., 2000; Nielsen & Molich, 1990a; Nielsen, 1992; Nielsen, 2000), it is emphasised how important it is to make sure that the participants and evaluators that are being used for the evaluations are part of a carefully selected group of users that match certain criteria concerning their expertise within the domain context of the product that is being evaluated. There are several arguments for why this should be ensured; some studies indicate that only participants with domain expertise should be used for this type of evaluations, as they are the only participants who would be able to be representative of actual end users (if of course there is a specified domain context), and that evaluation results would not be valid if they were not domain experts (Jenkins et al., 2003; Rubin & Chisnell, 2008, p. 115).

However, investigating the difference between how domain experts and novices yield data results in usability tests, is not something new. We have gone through examples of studies where this difference has been studied, and where the results vary, and what we found was, that even though some of this literature emphasises on the importance of domain expertise in participants for evaluations processes, many studies were conducted trying to demonstrate *why* this is, and how data results are affected if the domain expertise is not taken into consideration and is a part of the research design. However,

some of these studies (for example Karapanos et al., 2008 and Kjeldskov et al., 2010) concluded results that show the contrary: Domain expertise was found not to be the most crucial factor in evaluation processes, as the results from domain expert participants in many cases were as useful, or even less useful, than the results from domain novice participants.

We have not found any studies that deny the fact that domain expertise matters in evaluations, but as Jakob Nielsen described in his 2000 article, the focus on when to use participants with or without relevant domain expertise has shifted during the last few decades, and continues to shift because of how quickly different systems and products evolve and change, both in actual design, but also for *who* they were designed (Nielsen, 2000).

So, what does it all matter? Why should you ever use participants with little or no domain expertise, if the product or system that is being evaluated has one or more clear target groups or domain contexts that are important to how it works and who it is being targeted at?

We have defined the term *usability* by looking at which attributes the term is often being constructed from, and have split the term into parts based on the different attributes, so that we could investigate which usability attributes that are more affected by domain expertise, and how the test results differ for each of these attributes, rather than making observations on behalf of usability evaluations as a whole. The reason for doing so, is that we do not think that each of the attributes are affected equally by domain expertise, and that conducting usability evaluations is a process that is dependent on the type of product you are evaluating, and what parts of that product you are testing. What we discovered is that this understanding is right. The various attributes, we found, were affected differently.

One of the tendencies we found, was that test participants with little or no domain expertise relevant to the domain context of the product being tested, unsurprisingly yielded results that were a little more unfocused and scattered, compared to the results gathered in similar tests with domain expert test participants. We saw this when studying how domain expertise influences the way that attributes such as effectiveness, usefulness (in solving tasks), and user satisfaction, where the novice participants' test results are not as fruitful as the results from expert participants. The reason for this, we have seen, is that the experts are able to better justify in which areas the websites are lacking relevant information, or when a specific functionality is missing (or redundant), likely because their mental models of these types of websites are more developed and can be used to specify important issues in better detail. They are able to draw on some of their previously gained knowledge, which helps them relate usability issues to situations and contexts in which the same issue might not have

been a problem, and better express how they would solve or circumvent the same type of issue in this new context (Tullis & Albert, 2013, pp. 100-101). The results from the participant group of domain novices is still useful, however, but not as specific and precise, which in some cases would be undesirable.

In other attributes, we discovered how domain expertise is not playing a particular important role in terms of affecting evaluation results. For example when we compare the results of both participant groups number of clicks in tasks where they had to find specific information, as an incentive to measure the websites' effectiveness. In average, the expert participants spent less clicks finding the right information, but only marginally. The novices did however spend quite a bit more time solving the tasks and finding the information, as the combined average time (for all Think-Aloud tasks) spent for the novice participants was 725,8 seconds, and only 586,8 seconds for the experts (which means the experts were 23,7% faster at solving the tasks, than the novices in average).

The question is, if being the slower target group is equal to yielding test results that are of inferior quality? When we evaluated the learnability attribute, we found that having participants spend more time solving the tasks is not a bad thing, as it opens up for more "discussion" in the data (meaning that the participants had more time to express their thoughts and reasoning for their actions). On the other hand, having a participant group that is able to solve the tasks faster overall, is great for comparison – especially for an attribute such as learnability that focuses on how well users learn to utilise the system to their advantage (in terms of solving tasks and reaching goals). This is good, because it allows us to compare the results and see *why* participants with domain expertise were able to solve the tasks quicker. When doing so, we learned that it might have been because the expert participants had their own methods they used to find specific information, which actually provoked new design ideas for how the content of the websites could be made manageable to search and find.

In other attributes, such as navigation and labelling, we saw a tendency of how domain expertise is useful for evaluating Information Architecture that is targeted at users with the *same* type of domain expertise. However, we also saw, that if the product is being targeted at users with the same or *other* types of domain expertise, just as Feriecenter Slettestrand's main website is, using a participant group where the participants only have expertise within *one* of these domains, their test results are not necessarily as useful, as the participant group with no particular expertise. This could indicate, that participants with domain expertise within the same field, have a tendency to agree on the same principles, which we demonstrated in the section on labelling, where we saw that the domain expert

participants seemed to agree more on how they would sort their cards in the Card Sorting, so that it would make most sense to them. The same principle was discovered in the section where we investigated how the results differed when evaluating navigation. Here, the expert participants' results are also more precise and agreed upon with other expert participants, but since the context is not bound to *the* single domain that they are experts in, but rather a context of several domains, their results might actually not be as useful without another participant group's results to compare them to.

In reality, the best action might actually be to not limit evaluation participants to a single group of participants with the *same* type of domain knowledge, but rather try to expand the number of different perspectives, by including participants with either none or dissimilar types of domain expertise, so that the *multidimensional puzzle* can be solved, which would provide a common ground for end users to be able to utilise the product most efficiently and effectively, but independently and based on their subjective experiences and personal flexibility within those experiences.

5.26 The Number of Participants

We have already argued for how we chose the number of participants for the two user test methods, and how the numbers were based on earlier studies where other researchers found those numbers to be sufficient (around five participants for a Think-Aloud test, and around 10 for Card Sortings are argued to be necessary to find all the most critical issues, but not so many that the results become too redundant).

However, the question is if these numbers (even though we doubled them, because we had two participant groups) are enough to draw conclusions on how domain expertise truly influences test results, or if it requires a larger number so that we could measure it in a more quantitative fashion?

It might have been useful to be able to compare our qualitative findings with a quantitative study focusing on the same problem, and it would be very interesting to see if our findings would compare to a quantitative study. This has not been an option for us though, because of various restrictions, but we do believe that our findings can be used to at least indicate some of the tendencies that would probably also be found in a quantitative study of the same type, as our test results have been quite consistent throughout the tests, and across the test methods.

5.27 Retrospective Thoughts on used Methods and Research Design

If we were to suggest three different areas in which we would have changed our research design to ensure even better test results, now that we have learned new aspects, the first suggestion would be

that it would have been useful to test our hypothesis on a product that had an ultra-specific domain, rather than a product with several, but equally important domains. The mountain biking domain of Feriecenter Slettestrand's websites is still an important domain, and includes aspects that domain novices might not understand or know much about, but it is still somewhat approachable by most people. Had the product been very domain specific, and only have focused on a single, esoteric domain, some of the test results and differences between the two participant groups might have become even more clear.

The second suggestion would be to test over a longer time period. Some of the studies we used to investigate behaviour difference in novices and experts were longitudinal (for example the studies by Karapanos et al. (2008) and Kjeldskov et al., (2010)), meaning they were conducted over long time periods, which makes it easier to see how participants that started as novices changed their behaviour when they had become experts, and how that would influence the test results. This would eliminate elements such as random behaviour variance in tests like ours, as it would use the same participants at least twice.

The third suggestion would be to fine-tune the research design to also include tasks that induced "errors", for evaluating the usability attribute that is concerned with errors and safety. We did not do so, which meant that we discovered that the attribute's test results were nearly impossible to compare, as errors happened by random chance, but not enough for us to draw a conclusion or find any very important tendencies that could help us investigate how domain expertise influences that specific usability attribute.

Worth mentioning, is also the Card Sorting method. Being somewhat low-tech, it provides a basic, but very useful, approach to Information Architecture processes (such as the website structure, decide what to put on websites, labelling and navigation) that is not normally directly connected to usability studies (Petrie et al., 2011; U.S. Department of Health and Human Services, 2015a). Since we have expounded on our understanding of how Information Architecture can contribute to the understanding of usability, we decided to use it to further understand how its data can be utilised in usability evaluations, as it allows for a deeper understanding of some of the usability attributes. Even though the method is not traditionally connected directly to usability, we think that it should at least be considered in other usability evaluations, as our opinions is that it can provide information which can be used to support Think-Aloud test data in evaluations, especially since the method is so versatile (low-tech, easy to setup and can support many tasks).

6.0 Conclusion

The primary goal of this thesis has been to investigate how participant domain expertise affects data results, which are gathered during evaluation processes, and how domain expertise should be considered by researchers within the field of usability and Information Architecture, when framing a research design that incorporates test participants.

This goal originated from the fact that, when examining the methods on how to conduct usability and Information Architecture evaluations, the general opinion is that when recruiting participants to use in the evaluation process, the emphasis is that it is necessary to use participants with domain expertise within the domain contexts of the product, system or website being evaluated. Why this is, is not always clear, which is why we set up a research design that would inquire into this problem, and help us investigate how the results could be relevant for evaluators and evaluation processes in the future.

We found that participant domain expertise does in fact influences the way participants reflect on and complete context-relevant tasks, both when the tasks are relevant to their domain expertise, but also when they are not. Our results are divided into the different attributes of the usability term, as we consider that term to change meaning and attributes dependent on how and when it is being used. The attributes we used to define the term are based on several definitions, and have been employed in collaboration with attributes from the Information Architecture discipline, as it contains aspects such as navigation and labelling that helps shape and create the system or website in a way that supports good usability. Our definition of usability ended up being divided into attributes consisting of:

1. Effectiveness, usefulness and efficiency
2. Learnability and satisfaction
3. Errors and safety
4. Labelling
5. Navigation

When comparing the differences between the results from participants with domain expertise, and participants that are domain novices, we found that when evaluating usability attributes such as effectiveness, usefulness, efficiency, learnability and satisfaction, generally domain expert participants were faster at solving the tasks they were given. However, we also found that solving tasks quickly is not necessarily a good thing in an evaluation context. In terms of the effectiveness, usefulness and efficiency evaluation, the novice participants spent more time solving the tasks and

completing goals, but that extra time spent was used to study and verbalise their thoughts in a manner that was more useful in an evaluation process, whereas the expert participants seemed keener on actually solving the tasks.

The same was found during the evaluation of learnability; the expert participants solved tasks faster, but having novice participants spend more time on the tasks made it easier to distinguish when learning was actually taking place (and where it was not). In terms of the evaluation of the satisfaction attribute, the experts were better at expressing their subjective perceptions of what types of content they thought was necessary on Feriecenter Slettestrand's websites, and what information they expected to find. Their demands were better defined and more precise, which was very helpful to understand how domain expertise is important, especially when evaluating the attribute of satisfaction, whereas the novice participants had a harder time relating to the domain, and therefore also what they would find necessary to heighten their satisfaction.

The expert participants were also more specific in terms of the labelling attribute, where they found it easier to relate their domain expertise to the actual content of the websites. We saw that there is also only little resemblance between the results of the two participant groups' data results, when they were asked which labels they agreed would be most important for a website front page, indicating that the novice participant's results are more scattered and less precise. The two participant groups do however sort into categories much alike. When looking at the most-used labels for headlines, as both groups used the same four headlines for both the main and mountain biking website. Therefore, when evaluating labelling, expertise on the relevant domain (and the relevant domain content) is an important factor, which makes domain expert participants more useful when evaluating this attribute.

In terms of the navigation attribute, domain context is also important. We found that expert participants are especially important when evaluating navigation that is directly related to the domain that they are experts in. However, when evaluating navigational aspects that are not directly connected to a specific domain, but rather several domains, the novice participants yielded data results that were as good, if not better than what the expert participants provided. Considering these aspects, for the navigation attribute, we actually found the best evaluation results when combining the data from both participant groups, showing that only using *one* participant type for evaluating this attribute might not be the best choice, but that the results of the evaluation can be more diverse and useful overall if several participant types are used.

For the last attribute, errors and safety, we did not find a direct connection between participants' domain expertise and how it affects the data results, which was caused by inadequate research design. However, for this attribute, domain expertise might not be as important as technical expertise, as this attribute focuses more on technical skills, rather than how well the participants relate to a specific domain.

The findings from each of these attributes, and the way that we formed our research design, has led to a number of suggestions for how we would recommend using domain expert and/or novice participants for the various attributes of usability and Information Architecture evaluations. These recommendations are summed up in figure 36 on the next page.

Our results have been continuously compared to the results from other studies that have investigated how domain expertise influenced usability data results (Bednarik & Tukiainen, 2005; Botella et al., 2014; Dou et al., 2009; Karapanos et al., 2008; Kinney et al., 2008; Kjeldskov et al., 2010; Lazonder et al., 2000; Nielsen & Molich, 1990a; Nielsen, 1992; Nielsen, 2000), and have shown that it is not always an advantage to only use domain experts, as proposed by for example Jenkins et al. (2003) and Rubin and Chisnell (2008). Unlike those studies, we found that domain novices in some cases yield even better data than domain experts, which is important when framing a research design, as it can help to save valuable time and other resources and still contribute data results that are valuable in evaluation contexts.

However, since these findings are based on our data from this single thesis, we would recommend other evaluators to experiment with how participant domain expertise affects other usability and Information Architecture evaluation results, as it would require additional studies on this problem to determine exactly when and why to use one participant group over another. Our results should therefore be considered a contribution for other researchers to use, when they are designing their own research design for usability and Information Architecture evaluations. The results help to show how one might expect evaluation results to be affected, in contexts that are comparable to the context of this thesis. Having other studies confirm or deny our findings, or at least compare our results with their own, would lead to better arguments as to *why* it is important to include domain expert participants in the research design of usability evaluations, rather than it just being a single-sided view where it seems like you should always only use expert participants.

	Domain Experts	Domain Novices	A combination
Usefulness, effectiveness and efficiency	The domain expert participants solved tasks considerably faster, but used around the same number of clicks as novices to do so.	The domain novices were slower, but considering the research method (Think-Aloud), being the slower group is not necessarily bad, but can provide for data that is generally better and more diverse.	It depends on what you are evaluating. If there are specific design choices that needs evaluation, experts might be the better choice. For a more general evaluation, novices seem to discover more issues.
Learnability	Expert participants solve tasks faster than novices, which can make it harder to analyse the data.	Novice participants solve tasks considerably slower than experts, but this might be useful for distinguishing when learnability is a factor.	A combination of the two participant groups could help deploy solutions that would fit users that are somewhere in between the two groups' domain expertise.
Satisfaction	Expert's demands for what they find subjectively satisfying is more precise and better defined.	Novices have a harder time relating to foreign domains, and so satisfaction can be hard to measure.	Satisfaction is a very subjective usability attribute, and it is up to the evaluator to choose what type of participants he would find most effective.
Errors and Safety	Insufficient data, but technical expertise might be more useful, than domain expertise.	Insufficient data, but technical expertise might be more useful, than domain expertise.	Insufficient data, but technical expertise might be more useful, than domain expertise.
Labelling	Evaluating the navigational labels gave the most precise results in the group consisting of experts, especially for labels used on the front page.	The novice group gave results resembling the experts, but their results are scattered and not as precise as the experts' results.	If it is difficult to find participants fitting the user group requirements, novices can give useful results concerning navigational labels.
Navigation (Domain specific)	Evaluating the navigation of a domain specific website, you will need to use participants with relevant domain expertise, as the results from the experts give the best solution.	Using novices for evaluating the navigation of a website that are directed at a certain domain specific area will not be useful, as the results will not resemblance the results that participants with domain knowledge yield. This will lead to a navigational scheme not relevant to intended users.	This will most likely lead to a mixed result that can be difficult to evaluate.
Navigation (General/non-domain specific)	Only choosing participants with domain expertise on one part of a website can lead to results that only focus on the one domain the users are interested in.	The novices group may not have domain knowledge on one of the content areas of a multi domain website, but they can still give results that refer to the general navigation of the website.	A combination of participants with and without domain expertise can be relevant, if the given website has multiple domain specific content areas, as the experts focus at their own domain, rather than on the general content.

Figure 36 – Summary of how domain expertise affected data yielded from our test method results.

7.0 References

1. Alves, D. (2011). *How to Design the Best Navigation Bar for Your Website*. December 8, 2011. http://mashable.com/2011/12/08/design-navigation-bar/
2. Anderson, R. E., Johnson, D. G., Gotterbarn, D. & Perrolle, J. (1993) *Using the New ACM Code of Ethics in Decision Making*. February vol. 36, no 2. Communications of the ACM.
3. Barnum, C. M. (2011). *Usability Testing Essentials – Ready, Set… Test!*. Morgan Kaufmann.
4. Bednarik, R., Tukiainen, M. (2005). *Effects of Display Blurring on the Behavior of Novices and Experts during Program Debugging*. In CHI '05 extended abstracts on Human factors in computing systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005, ACM Press, pp. 1204-1207.
5. Benyon, D. (2010). *Designing Interactive Systems – A comprehensive guide to HCI and interaction design* (2nd Ed.). Essex, Pearson Education Limited.
6. Berry, D. C. & Broadbent, D. E. (1990). *The Role of Instruction and Verbalization in Improving Performance on Complex Search Tasks*. Behaviour & Information Technology 9, 3. May-June, 175-190.
7. Berry, D.C. & Broadbent, D.E. (1990). *Role of Instruction and Verbalization in Improving Performance on Complex Search Tasks*. Behaviour & Information Technology 9, 3 (May-June), 175-190.
8. Bødker, K., Kensing, F. & Simonsen, J. (2008). *Professionel IT-Forundersøgelse – Grundlag for Brugerdrevet Innovation*. Forlaget Samfundslitteratur.
9. Botella, F., Alarcon, E., & Peñalver, A. (2014). *How to classify experts in usability evaluation*. Proceedings of the XV International Conference on Human Computer Interaction.
10. Budd, A. (2007). *Heuristics for Modern Web Application Development*. Blogography. January 17, 2007. Available at: http://www.andybudd.com/archives/2007/01/heuristics_for_modern_web_application_development/
11. Bussolon, S., Russi, B. & Missier, F. (2006) *Online Card Sorting: As Good as the Paper Version*. Proceedings of the 13th European Conference on Cognitive Ergonomics: Trust and Control in Complex Socio-Technical System (ECCE). ACM Press.
12. Clifton, B. (2012). *Advanced Web Metrics with Google Analytics* (3rd ed.). John Wiley & Sons.
13. Danmarks Statistik. (2014). *It-anvendelse i befolkningen*. Retrieved from: http://www.dst.dk/pukora/epub/upload/18686/itbef.pdf

14. Denning, S., Hoiern, D., Simpson, M. & Sullivan, K. (1990). *The Value of Thinking-Aloud Protocols in Industry: A Case Study at Microsoft Corporation*. Proceeding. Human Factors Society 34th Annual Meeting, 1285-1289.
15. Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R., Chang, R. (2009) *Comparing Usage Patterns of Domain Experts and Novices in Visual Analytical Tasks*. In *Sensemaking Workshop*, ACM CHI 2009.
16. Elling, S., Lentz, L. & de Jong, M. (2011). *Retrospective Think-Aloud Method: Using Eye Movements as and Extra Cue for Participants' Verbalization*. CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Pages 1161-1170. ACM New York, NY, USA.
17. Ericsson K. A. & Simon, H. A. (1993) *Protocol Analysis – Verbal Reports as Data*. A Bradford Book. The MIT Press. Cambridge, Massachusetts London, England.
18. Freeman, B. (2011). *Triggered Think-Aloud Protocol: Using Eye Tracking to Improve Usability Test Moderation*. CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
19. Guan, Z., Lee, Z., Cuddihy, E. & Ramey, J. (2006) *The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking*. CHI, April 22-27. ACM Montréal, Québec, Canada.
20. Harper, M. E, Jentsch, F., van Duyne, L. R., Smith-Jentsch, K. & Sanchez, A.D. (2002) *Computerized Card Sort Training Tool: Is It Comparable to Manual Card Sorting?* Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting, 2049-2053.
21. Hornbæk, K. (2006). *Current Practice in Measuring Usability: Challenges to Usability Studies and Research*. *International Journal of Human-Computer Studies*, 64 (2), pp. 79-102.
22. Instone, K. (1997) *Site Usability Heuristics for the Web*. Web Review, October. Available at: http://instone.org/heuristics
23. ISO/IEC. (1998). *ISO 9241-11:1998 - Ergonomic requirements for office work visual display terminals (VDTs) – Part 11: Guidance on usability*. London: BSI.
24. Jenkins, C., Corritore, C. L. & Wiedenbeck, S. (2003). *Patterns of Information Seeking on the Web: A Qualitative Study of Domain Expertise and Web Expertise*. IT & Society, Volume 1, Issue 3, Winter 2003, pp. 64-89.
25. Kalbach, J. (2007). *Designing Web Navigation – Optimizing the User Experience*. O'Reilly Media, Sebastopol, CA.

26. Karapanos, E., Hazzenzahl, M. & Martens, J.B. (2008). *User Experience Over Time*. CHI 2008 Proceedings. ACM 978-1-60558-012-8/08/04.
27. Kinney, K. A, Huffman, S. B. & Zhai, J. (2008). How Evaluator Domain Expertise Affects Search Result Relevance Judgments. Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 591-598.
28. Kjeldskov, J., Skov, M. B., & Stage, J. (2010). *A Longitudinal Study of Usability in Health Care: Does Time Heal?*. *International Journal of Medical Informatics*, *79*(6), e135-e143.10.1016/j.ijmedinf.2008.07.008
29. Lazonder, A. W., Biemans, H. J.A. & Wopereis, I. G.J.H. (2000). *Differences between novice and experienced users in searching information on the World Wide Web*. J. Am. Soc. Inf. Sci., 51: 576–581. doi: 10.1002/(SICI)1097-4571(2000)51:6<576::AID-ASI9>3.0.CO;2-7
30. Lewis, C.H. (1982). *Using the "Thinking Aloud" Method in Cognitive Interface Design*. IBM Res. Rep. RC-9265. Yorktown Heights, N.Y.
31. Maurer, D. & Warfel, T. (2004) *Card Sorting: A Definitive Guide*. Boxes and Arrows. April 2004. http://boxesandarrows.com/card-sorting-a-definitive-guide/
32. Molich, R., Ede, M. R., Kaasgaard, K. & Karyukin, B. (2004). *Comparative Usability Evaluation*. Behaviour & Information Technology, 23:1, pp. 65-74, DOI: 10.1080/0144929032000173951.
33. Morville, P. & Rosenfeld, L. (2006*). Information Architecture for the World Wide Web* (3rd Ed.). O'Reilly Media, Sebastopol, CA.
34. Nielsen, J. & Landauer, T. K. (1993) *A Mathematical Model of the Finding of Usability Problems*. Proceedings. ACM INTERCHI '93 Conference. Amsterdam, The Netherlands, 24-29 April, pp. 206-213.
35. Nielsen, J. & Molich, R. (1990a). *Heuristic Evaluation of User Interfaces*. CHI '90 Proceedings. April 1990.
36. Nielsen, J. & Molich, R. (1990b). *Improving a Human-Computer Dialogue*. Communication of the ACM. March 1990, Volume 33, Number 3.
37. Nielsen, J. (1992). *Finding Usability Problems through Heuristic Evaluation*. Proceedings ACM CHI'92 Conference. Pages 373-380.
38. Nielsen, J. (1993). *Usability Engineering*. Academic Press, Boston, MA.
39. Nielsen, J. (1995). *Usability Testing for the 1995 Sun Microsystems' Website*. May 25, 1995. http://www.nngroup.com/articles/usability-testing-1995-sun-microsystems-website/

40. Nielsen, J. (2000). *Novice vs. Expert Users*. Alertbox, February 6, 2000. http://www.useit.com/alertbox/20000206.html
41. Nielsen, J. (2004). *Card Sorting: How Many Users to Test*. July 19, 2004. http://www.nngroup.com/articles/card-sorting-how-many-users-to-test/
42. Nielsen, J. (2012). *Thinking Aloud: The #1 Usability Tool*. Januar 6, 2012. http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/
43. Olsen, A., Smolentzow, L. & Strandvall, T. (2010) *Comparing Different Eye Tracking Cues When Using the Retrospective Think Aloud Method in Usability Testing*. Proceeding, BCS '10 Proceedings of the 24th BCS Interaction Specialist Group Conference. Pages 45-53.
44. Petrie, H. & Precious, J. (2010). *Measuring User Experience of websites: Think aloud protocols and an emotion word prompt list*. CHI 2010: Work-in-Progress (Spotlight on Posters Days 1 & 2). April 12–13, 2010, Atlanta, GA, USA
45. Petrie, H., & Power, C. (2012). *What Do Users Really Care About? A Comparison of Usability Problems Found by Users and Experts on Highly Interactive Websites*. CHI '12, May 5–10, 2012, Austin, TX.
46. Petrie, H., Power, C., Cairns, P. & Seneler, C. (2011). *Using Card Sorts for Understanding Website Information Architectures: Technological, Methodological and Cultural Issues*. Human-Computer Interaction – INTERACT 2011, Lecture Notes in Computer Science Volume 6949, 2011, pp 309-322.
47. Reiss, J. (2012). *Usable Usability – Simple Steps for Making Stuff Better*. John Wiley & Sons.
48. Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction Design: Beyond Human-Computer Interaction* (3rd Ed.). Chichester, Wiley.
49. Rubin, J. & Chisnell, D. (2008). *Handbook of Usability Testing (2nd ed.) – How to Plan, Design, and Conduct Effective Tests*. Wiley Publishing. Indianapolis, IN.
50. Russell-Rose, T. & Tate, T. (2013). *Designing the Search Experience – The Information Architecture of Discovery*. Morgan Kaufmann, Waltham, MA.
51. Snitker, T. V. (2001). *Brug Brugerne – Og Skab Mere Brugervenlige Web-sites*. Ingeniøren|bøger.
52. Sova, D. H. & Nielsen, J. (2003). *234 Tips and Tricks for Recruiting Users as Participants in Usability Studies*. Nielsen Norman Group. http://www.nngroup.com/reports/how-to-recruit-participants-usability-studies/

53. Spencer, D. (2009) *Card Sorting: Designing Usable Categories*. Brooklyn, NY. Rosenfeld Media.
54. Tidwell, J. (2011). *Designing Interfaces – Patterns for Effective Interaction Design* (2nd ed.). O'Reilly Media, Sebastopol, CA
55. Tullis, T. & Albert, B. (2013). *Measuring the User Experience – Collecting, Analyzing, and Presenting Usability Metrics*. Second Edition. Morgan Kaufmann, Waltham, MA.
56. U.S. Department of Health and Human Services. (2015a). *Card Sorting.* Washington, DC. Retrieved from: http://www.usability.gov/how-to-and-tools/methods/card-sorting.html
57. U.S. Department of Health and Human Services. (2015b). *Usability Evaluation Basics.* Washington, DC. Retrieved from: http://www.usability.gov/what-and-why/usability-evaluation.html
58. Watson, J. B. (1920). *Is Thinking Merely the Action of Language Mechanism*. British Journal of Psychology, 11, 87-104.
59. White, R. W., Dumais, S. T. & Teevan, J. (2009). *Characterizing the Influence of Domain Expertise on Web Search Behavior*. Microsoft Research. Proceedings of Web Search and Data Mining (WSDM) 2009.
60. Wills, J. & Hurley, K. (2012). *Testing Usability and Measuring Task-Completion Time in Hospital-Based Health Information Systems: A Systematic Review*. 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE).
61. Wright, R. B. & Converse, S. A. (1992). *Method bias and concurrent verbal protocol in software usability testing*. Proceeding. Human Factors Society 36th Annual Meeting, 1220-1224.