COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DANMARK

www.cbs.dk

ISSN 0906-6934
ISBN 978-87-593-8443-5

ISBN 978-87-593-8443-5
9 788759 384435

An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

PhD Series 30.2010

Copenhagen
Business School
HANDELSHØJSKOLEN

An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

Qingxin Shi

LIMAC PhD School
Programme in Informatics

PhD Series 30.2010

An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

1st edition 2010
PhD Series 30.2010

ISBN: 978-87-593-8443-5
ISSN: 0906-6934

LIMAC PhD School is a cross disciplinary PhD School connected to research communities within the areas of Languages, Law, Informatics, Operations Management, Accounting, Communication and Cultural Studies.

An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

Qingxin Shi

LIMAC PhD School, Programme in Informatics
Copenhagen Business School
Department of Informatics

Acknowledgement

First, I would like to give my sincere thanks to my supervisor, Torkil Clemmensen. Without his kind help and encouragement in the past three years, I do not think I could have ever finished my PhD studies. Being a foreign student who had never studied abroad before, I had many new experiences, and during this time Torkil Clemmensen always had patience in answering my questions and giving me valuable advice on how to carry out this PhD project. He also did his best to help me to apply for funding for recruiting participants and finding usability practitioners in this research. When I was writing this thesis, I received invaluable feedback from him.

It is hard for me to express enough gratitude to my co-supervisor, Kan Zhang, at the Institute of Psychology, Chinese Academy of Sciences. His profound knowledge and generosity inspired me to carry out the PhD research in Denmark. I thank Xianghong Sun, also from the Institute of Psychology, Chinese Academy of Sciences, for her support of the data collection and data analysis. I also thank Chenfu Cui for his effort in coding the videos.

Many thanks are given to the usability consulting companies of “Snitker & Co.” and “Userminds” for their willingness to take part in this PhD project. Many usability specialists in their companies attended my study as evaluators, and my thanks go to “Snitker & Co.” for helping me recruit users. I would like to thank all the usability specialists who took part in my study. Without them, this research would not have been possible. I thank Morten Hertzum and Ravi Vatrapu for their valuable comments on the whole thesis in the pre-defence. I thank Jacob Nørbjerg and Volker Mahnke for their advice on how to write the introduction and discussion chapters.

I am grateful for the good atmosphere and studying conditions that the Department of Informatics provided. Thanks to Anni Olesen, Martin Tong and Tasja Rodian for making my work here easier.

My thanks also go to Saihong Li, who has supported me throughout my doctorate research with friendship and generosity.

My warm thanks are given to my husband, Kerong Chen, for his support and understanding of the PhD studies in the past three years. I also thank all my friends and family for their spiritual support at every stage of my work in Denmark.

Abstract

Usability evaluation methods are widely used to assess and improve the user interface design. This dissertation investigates the thinking aloud usability testing from a cultural perspective. In a test situation, representative users are required to verbalize their thoughts as they perform their tasks while using the system, and an evaluator observes the user's task performance and comes up with usability problems. The primary goal of a usability test is to find a list of usability problems.

In this research, the impacts of evaluators' and users' cultural backgrounds on both the result and the process of the thinking aloud usability testing were investigated. Regarding the results of the usability testing, the identified usability problem was the main focus, whereas for the process of testing, the communication between users and evaluators was the main focus.

In this dissertation, culture was regarded as cognitive styles and communication orientations. For the theories of thinking aloud, both Ericsson and Simon's classic model, and Boren and Ramey's revised model for usability testing were taken into account. Based on the culture theories and thinking aloud models, hypotheses were developed to investigate the evaluators' identified usability problems in different cultural settings, and themes were put forward to investigate the evaluators' and users' communications.

In order to investigate the hypotheses and themes, an experimental study was conducted. The experimental design consisted of four independent groups with evaluators and users from similar or different cultures (Danish and Chinese). Empirical data were collected by using background questionnaires, usability problem forms, usability problem lists, video recordings of the testing and interviews. The usability testing software "Morae" was used to record the whole testing, including the faces of the evaluators and users, the screen and keyboard activities. Evaluators' and users' communications were analyzed by the behavioural coding and analysis software "Observer XT 8.0" with a well defined coding system.

The results of the systematic study of the thinking aloud usability testing in the context of the intra- and inter-cultural usability engineering show that the evaluators' cultural backgrounds do have some influences on the usability testing; however, the influences are different for the tests with Western and East Asian users. The main findings of this research have implications for both usability research and practice. The methodological approach also gives inspiration for usability evaluation studies.

Dansk Resume

Usability evalueringsmetoder er almindeligt anvendt til at vurdere og forbedre designet af brugergrænseflader. Denne afhandling undersøger ”tænke-højt usability testen” fra et kulturelt perspektiv. I en tænke-højt usability test situation skal repræsentative brugere sætte ord på deres tanker, når de udfører deres opgaver, og mens de bruger systemet. En evaluator observerer brugerens opgaveløsning og beskriver usability problemerne. Det primære mål for en usability test er at finde en liste over usability problemer.

I dette forskningsprojekt var fokus på hvordan evaluatorernes og brugernes kulturelle baggrunde påvirkede både resultatet og processen ved tænke højt usability test. Med hensyn til resultaterne af usability tests var fokus på studiet af de identificerede usability problemer, mens fokus ved studiet af test-processen var på kommunikationen mellem brugere og evaluatorer.

I denne afhandling blev kultur betragtet som kognitive stile og som måder at kommunikere på. Vedrørende teorier om tænke-højt metoden blev Ericsson og Simon's klassiske model, og den af Borén og Ramey reviderede model for usability tests, taget i betragtning i denne afhandling. Baseret på kultur teorierne og tænk-højt modellerne, blev hypoteser udviklet til at undersøge evaluatorernes identifikation af usability problemerne i forskellige kulturelle miljøer, og temaer blev fremført med henblik på at undersøge evaluatorerne og brugernes kommunikation.

For at undersøge hypoteserne og temaerne blev en eksperimentel undersøgelse udført. Det eksperimentelle design bestod af fire uafhængige grupper med evaluatorer og brugere fra de samme eller forskellige kulturer (dansk og kinesisk). Empiriske data blev indsamlet ved hjælp af baggrunds-spørgeskemaer, usability problem beskrivelses-formularer, usability problem-lister, og videooptagelser af test og interviews. Usability test software, "Morae", blev brugt til at registrere hele testforløbet, herunder evaluatorers og brugeres ansigter, og deres skærm og tastatur aktiviteter. Evaluatorernes og brugernes meddelelser blev analyseret ved hjælp af kodning i adfærdsanalyse-software, "Observer XT 8.0", med et veldefineret kodesystem.

Resultaterne af den systematiske undersøgelse af tænke-højt usability tests i forbindelse med intra- og interkulturelle usability teknikker viser, at evaluatorernes kulturelle baggrunde i nogen grad påvirker usability tests, men påvirkningerne er forskellige for tests med vestlige og østasiatiske brugere. Resultater af nærværende forskning har konsekvenser for både usability forskning og praksis. Den metodiske tilgang kan give inspiration til nye usability evalueringsundersøgelser.

Table of Contents

List of Tables

List of Figures

1 Introduction

Usability evaluation is an important element of systems development (Hertzum, Hansen, & Andersen, 2009; Hornbæk, 2009). With the rapid development of science and technology, the functions of the products become more and more complex. In order to ensure that the products will be accepted by users, it is important for the products to have a good usability (Benbunan-Fich, 2001; Karsh, 2004; Nielsen, 1993). Therefore, it is necessary to have an effective usability evaluation method which can be used to assess and improve the products' usability.

Nowadays, culture is playing a more and more important role in the global market. There have been numerous studies on cultural influence on design and usability in the past few years (Aryee, Luk, & Fields, 1999; Bilal & Bachir, 2007; Bourges-Waldegg & Scrivener, 1998; Bourges-Waldegg & Scrivener, 2000; Carey, 1998; Day, 1998; De Angeli, Athavankar, Joshi, Coventry, & Johnson, 2004; Downey, Wentling, Wentling, & Wadsworth, 2005; Duncker, 2002; Fang & Rau, 2003; Griffith, 1998; Honold, 2000; Onibere, Morgan, Busang, & Mpoeleng, 2001; Rose & Zuhlke, 2005; Sacher, Tng, & Loudon, 2001; Shen, Woolley, & Prior, 2006; Smith, Dunckley, French, Minocha, & Chang, 2004; Sun, 2004). An increasing number of researchers realize that culture influences not only the systems' usability and design, but also the evaluation methods, such as focus group, structured interview and usability testing (Lee & Lee, 2007; Vatrapu & Pérez-Quiñones, 2006; Yeo, 1998a).

Usability evaluation methods (UEMs) are the methods or techniques used to perform evaluation of an interface design at any stage of its development (Hartson, Andre, & Williges, 2001). The goal of a usability evaluation method (UEM) is to "produce descriptions of usability problems observed or detected in the interaction design for analysis and redesign" (Hartson et al., 2001, p. 379). Compared to the studies of cultural influence on design and usability, studies on cultural influence on UEMs are relatively fewer (Clemmensen, 2006; Clemmensen & Goyal, 2005; Clemmensen et al., 2007; Lee & Lee, 2007; Vatrapu & Pérez-Quiñones, 2006; Yammiyavar, Clemmensen, & Kumar, 2008; Yeo, 1998a). Among all the UEMs, thinking aloud usability testing is regarded as the single most valuable usability engineering method for evaluating the usability of user interfaces (Nielsen, 1993, p 195). In this dissertation, the thinking aloud usability testing in intra- and inter- cultural settings is investigated.

1.1 Problem Statement

Thinking aloud is used as a usability evaluation method to gain insight into how users work with a product or interface. The thinking aloud usability testing has been "extensively applied in industry to evaluate a system's prototypes of different levels of fidelity" (Law & Hvanneberg, 2004, p. 9). In usability tests, representative users are required to verbalize their thoughts while performing pre-established tasks by using the system (Nielsen, 1993, p. 195). The primary goal of a usability test is to "derive a list of usability problems from evaluators' observations and analyses of users' verbal as well as non-verbal behavior" (Law & Hvanneberg, 2004, p. 9). Figure 1 shows the elements in a thinking aloud usability testing (Clemmensen, Hertzum, Hornbaek, Shi, & Yammiyavar, 2008; Clemmensen, Hertzum, Hornbaek, Shi, & Yammiyavar, 2009).

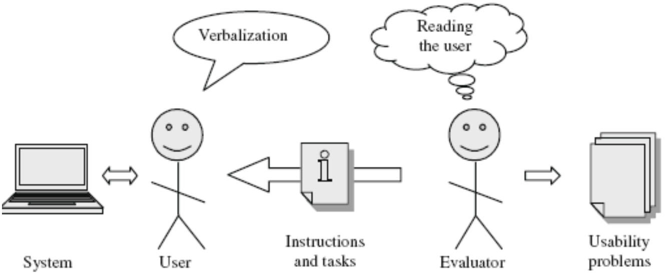


Figure 1: Reference model of thinking aloud usability testing (Clemmensen et al., 2009, p. 214)

In a thinking aloud usability testing, the user is the participant who interacts with the system and verbalizes his/her thoughts while doing the tasks. The evaluator is the usability professional who facilitates the testing, observes the user's task performing and comes up with usability problems. The tasks that the user conducts and the instructions that the user follows are given by the evaluator. Apart from presenting tasks and instructions, the evaluator also needs to "read the user" which means he/she has to observe the user's task performing behaviour and listen to the user's thought verbalization in order to understand not only the bad and good aspects of the system (Nielsen, 1993), but also to achieve the goal of the usability testing---finding usability problems (Hartson et al., 2001; Kaminsky, 1992; Nielsen, 1993).

Compared to other UEMs, usability testing is often regarded as an objective evaluation method since it involves the users' task performing behaviour. It has been used to verify the UEMs which only involve the users' or evaluators' opinions, such as in interviews, surveys and heuristic evaluations (Hvannberg, Law, & Lárusdóttir, 2007; Yeo, 2001b). The data from interviews and surveys are normally based on the participants' own judgment, whereas the data from heuristic evaluation are normally based on certain rules which are used to predict potential usability problems by evaluators. Usability problems from interviews, surveys or heuristic evaluations may not be real problems that the users may have when interacting with the interface/system (Hartson et al., 2001; Hvannberg et al., 2007; Yeo, 2001a). Many researchers regard usability testing as an objective method since it involves user performance (Hartson et al., 2001; Hvannberg et al., 2007; Yeo, 2001b). However, from Figure 1 we can see that the thinking aloud usability testing may not be as objective as researchers have commonly thought. Clemmensen, Hertzum et al. (2008, 2009) have discussed the impact of culture on thinking aloud usability testing based on the model in Figure 1. The authors point out that culture may influence the process of usability testing, such as the user's verbalization and the evaluator's reading of the user. Evaluators and users from different cultures may have different communication patterns in the testing. Moreover, evaluators with different cultural backgrounds may have different understandings of the target users' task behaviours which may result in extracting different usability problems.

In this research, the influence of culture on both the process and the result of the usability testing will be investigated. For the process of the usability testing, the focus will be on the communication between the user and the evaluator; for the result of the usability testing, the focus will be on the usability problems identified by the evaluators after the tests.

Even though the usability problem identification has been investigated by many researchers (Hertzum, 2006; Hornbaek & Frøkjær, 2008; Hornbæk & Frøkjær, 2005; Hvannberg et al., 2007; Keenan, Hartson, Kafura, & Schulman, 1999; Nielsen & Landauer, 1993; Virzi, Sokolov, & Karis), there are limited studies on the comparison of the usability problems in different cultural settings (Law & Hvanneberg, 2004; Vatrapu & Pérez-Quiñones, 2006). This study investigates the usability testing in different cultural settings and examines the impact of evaluators' and users' cultural backgrounds on the evaluators' usability problem finding and problem severity rating behaviours. The cultural settings in this research involve both intra- and inter-cultural usability testing settings.

As Nielsen (1993) indicates, the thinking aloud usability testing is a typical method used for formative evaluation. Formative evaluation focuses on identifying usability problems (Hartson et al., 2001, p. 375), not the performance of the tasks, and thus communication with users is necessary and important. However, how to communicate with users in order to find usability problems is not clear. In this research, both intra- and inter- cultural communications are involved. Intra-cultural communication is similarity-based, whereas inter-cultural communication is difference-based (Bennett, 1998). Intra-cultural communication is the type of communication that takes place between users and local evaluators. The communication between users and foreign evaluators could be regarded as "intercultural communication," which is an important topic involving anthropology, psychology, cultural studies and communication studies (Andersen, 2001; Gunykunst, 1993; Hall, 1989a, 1990; Hall & Hall, 1990; Kim & Gudykunst, 1988; Lull, 2001; Luzio, Gunthner, & Orletti, 2001; Samovar & Porter, 2003; Spencer-Oatey, 2000). Since communication "entails the idea of interdependence, a process having mutuality, shared activity, some form of linkage or connection with a message" (Sereno & Mortensen, 1970, p. 178), the communication between the local pairs may be different from that of the distant pairs (users with foreign evaluators). In this PhD thesis, the patterns or genres of the communication (Hall, 1989a; Hall & Hall, 1990; Kim & Gudykunst, 1988; Luzio et al., 2001; Yoshioka & Herman, 1999) for the evaluators and users with different cultural backgrounds are investigated.

Thinking Aloud Models

Two main thinking aloud models have been put forward by researchers. One is the classic thinking aloud model developed by Ericsson and Simon (1993) based on information processing theory. The other thinking aloud model is proposed by Boren and Ramey (2000) based on speech communication theory. In this thesis, both models are considered, but Boren and Ramey's model developed for the usability testing is the main concern (Clemmensen & Shi, 2008; Clemmensen et al., 2007; Dumas & Loring, 2008; Nørgaard & Hornbæk, 2006; Shi, 2008a; Tamler, 2003). In Boren and Ramey's thinking aloud model, "talk is not simply a form of action" performed by the user alone, "but a mode of interaction" between users and evaluators (Boren & Ramey, 2000, p. 267).

Culture Theories

In this dissertation, culture is regarded as different cognitive styles and communication orientations. Nisbett and his colleagues' work on different cognitive styles of Westerners and East Asians (Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002a; Nisbett, Peng, Choi, & Norenzayan, 2001) and Hall's work of the high- and low- contextual communication orientations (Hall, 1989a, 1990; Hall & Hall, 1990) are the main culture theories followed in this research. Nisbett and his colleagues' cultural psychology theory and empirical findings implicate that even though the evaluators are usability professionals who have the ability and skill to communicate with users and detect usability problems, if they are from different intellectual traditions from that of the users, the communication, usability problem finding and usability problem severity rating may still be influenced by the evaluators' local cultural perception and cognition. Nisbett's culture theory discusses the cognition and perception differences between Europeans, Americans vs. Chinese, Korean and Japanese (Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003). According to this theory, Europeans and Americans are named as Westerners, and Chinese, Korean and Japanese are named as East Asians (Nisbett, 2003). Westerners and East Asians who come from two different intellectual traditions have different cognitive styles. Supposedly, Westerners have an analytic cognitive style, whereas East Asians have a holistic cognitive style. The theory is very relevant to usability tests, since the thinking aloud process involves users' cognition and perception characteristics, and the results of the usability test, i.e., usability problems which are found by the evaluators, involve the evaluators' cognition and perception of the whole test process. Thus, Nisbett and his colleagues' studies on cognitive styles are imported into usability testing in this research.

Moreover, in the thinking aloud usability testing, evaluators' and users' high- and low-contextual communication orientations may influence both the usability problems and communication patterns. Hall regards culture as communication (Hall, 1990, p. 94). According to Hall's high- and low-context communication theory, a person from a high-context communication country may have problems when he/she is communicating with the person from a low-context communication country. On the one hand, if the communication patterns between the local pairs and distant pairs are the same, the usability problems may still be different, since they may have a different understanding of the information transmitted from the user/evaluator. On the other hand, if finding all the relevant usability problems is the main task for the local and foreign evaluators, the evaluators' and users' communication patterns may be changed due to those whom the evaluators/ users are with.

According to Nisbett's and Hall's culture theories, in this dissertation, Western culture includes countries with people who tend to have an analytic cognitive style and low-contextual communication orientation, and East Asian culture includes countries with people who tend to have a holistic cognitive style and high-contextual communication orientation.

1.2 Research Question

The main research question for this thesis is: To what extent does the evaluators' and users' cultural background influence the thinking aloud usability testing?

In this research, the cultural impact on thinking aloud usability testing is investigated. Evaluators and users with the same and different cultural backgrounds were invited to attend the study in order to examine the extent to which their cultural backgrounds impact the thinking aloud usability testing. As discussed in section 1.1, the research question is investigated from two perspectives—the result of the testing (usability problems) and the process of the testing (communications), presented as two sub-research questions. The evaluators' and users' cultural backgrounds may influence both usability problems and communications. In this research, Denmark is selected to represent the Western culture, and China is selected to represent the East Asian culture. Accordingly, the usability tests were conducted in Denmark and China with Danish and Chinese evaluators and users, respectively, in order to investigate the research question.

The research question divided into two sub-questions:

R1. In the thinking aloud usability testing, to what extent does the evaluators' and users' cultural background influence the evaluators' usability problem finding and usability problem severity rating?

In this study, both problem finding and problem severity rating behaviors are investigated. Identifying the usability problems in the user interface that "could result in human error, terminate the interaction, and lead to frustration on the part of the user" (Norman & Panizzi, 2006, p. 247) is one of the most important goals and purposes of the usability testing (Hartson et al., 2001; Law & Hvanneberg, 2004; Nielsen, 1993). The identified usability problems could be considered as the result of the usability testing and the performance of the communication between the evaluators and users. Prior empirical results strongly suggest that the cognitive styles and communication orientations vary across cultures (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2003; Nisbett, 2004; Nisbett & Norenzayan, 2002a), and thus there may be some misunderstandings between the evaluators and users with different cultural backgrounds. Western and East Asian evaluators may focus on different part of the culturally localized application which may result in finding different usability problems.

Furthermore, the usability problem severity rating may also be influenced by culture. Severity is an important "measure of the quality of each usability problem found by a UEM, offering a guide for practitioners in deciding which usability problems are most important to fix" (Hartson et al., 2001, p. 388). Hertzum and Jacobsen (2001) report that evaluators have different strategies for assessing the severity of usability problems. This study investigates whether the severity rating strategy varies across cultures, i.e., whether Western and East Asian evaluators have different preferences in rating the usability problem severity. Hence, the usability problem finding and the problem severity rating between the Western and East Asian evaluators will be examined in this research.

R2. To what extent does the evaluators' and users' cultural background influence the evaluators' and users' communications in the thinking aloud usability testing?

From the theoretical bases in this research, evaluators' and users' cultural backgrounds may have influence on their communication. The target of communication in the usability test is to transmit the intended message from the user to the evaluator and also from the evaluator to the user. As mentioned, evaluators and users with different cultural backgrounds may have different patterns of communication. Previous work shows that communication patterns can be at least partly attributed to cognitive styles (Littlemore, 2001). Based on Nisbett and his colleagues' research, people in the Western and East Asian countries tend to have different cognitive styles, thus the communication patterns may be different. Furthermore, according to Hall (Hall, 1989a; Hall & Hall, 1990), people in different cultures may have different communication orientations which may also show different patterns in their communication. In this research, the main focus is on investigating the verbal communication. A previous study (Yammiyavar et al., 2008) has investigated non-verbal communication, showing that there are some culturally specific gestures in thinking aloud usability testing. There may also be some specific verbal communication differences for the users and evaluators in different cultures. After analyzing the usability problems in sub-question 1, the communication will be investigated.

1.3 Motivations for this Study

Section 1.3 discusses the motivation for this study from the practitioner and theoretical point of view.

1.3.1 Motivation from the Practitioners' Point of View

With the advent of globalization and IT revolution, we can no longer overlook the aspect of culture in the design of user interfaces and products. Considering the cultural aspect has become one of the key factors for the success or failure of a global product. By accommodating more countries, multinational companies can earn more revenue from international markets (Yeo, 2001a). In order to capture the target market, the products or interfaces should be designed and tested for the target people. Rose and Zuhlke (2005) report that the earlier the localization is considered, the better effects and lower costs will be achieved over the usability life cycles. In a usability life cycle, one of the important phases is to evaluate and modify the product/ interface through iterative evaluation methods (Mayhew & Bias, 2005). However, previous studies have shown that culture influences the evaluation methods (Lee & Lee, 2007; Vatrapu & Pérez-Quiñones, 2006; Yeo, 1998a). The way of carrying out the thinking aloud usability testing in Western and East Asian countries may not be the same (Yeo, 1998a). This current research is conducted in order to have a better understanding of the thinking aloud usability testing in different countries. Through this study, we hope to give some valuable advice to usability practitioners on how to do thinking aloud usability testing in Western and East Asian countries.

Moreover, this research investigates not only local evaluators' usability problem finding, severity rating and communicating behaviors, but also investigates foreign evaluators' behaviors in the tests. Even though today the common approach to carrying out usability tests in a foreign country is to recruit local evaluators (Dray & Siegel, 2005), in some situations, foreign evaluators, instead of local ones, need to be used. The reasons are:

1) It may be more efficient to use foreign evaluators, especially in the situation of testing prototypes with target users in order to get quick feedback to the developers. For example, a research team in a company is trying to develop an interface and wants to know

whether it is good for people in different cultures. It may be good to consider the "discount usability engineering" (Nielsen, 1993, p. 17) which can be used to improve the interface in a fast and cheap way by using their own evaluators to do the tests with different users. Before an interface is finally implemented, it needs many evaluation life cycles (Mayhew & Bias, 2005). Usability evaluation is necessary in every application developing phase. If training the local evaluators, the company needs to do many things before the test, such as introducing how to do the usability test, the method and the product, clarifying the research purpose, the focuses of the products and so on. The local evaluator also needs to learn the video of a related test, conduct a mock session and a dry run (Dray & Siegel, 2005, P. 209). In order to modify the prototype, the developers in the application developing country need to wait for the final reports from the local evaluators, which can also take some time (Dray & Siegel, 2005; Dumas & Redish, 1999). If using the foreign evaluators who are from the application developing country, they need to communicate with the developers frequently while they are doing the tests. Thus, using foreign evaluators is sometimes more efficient than using local ones, especially when foreign evaluators know the project well and have done the tests in their own country.

2) It may be less costly to use foreign evaluators in some situations. Nowadays, professional evaluators are a valuable and expensive human resource. Further, many prototypes are not tested before they are produced as the final applications in the market, but need to be designed and redesigned through usability engineering lifecycle (Mayhew & Bias, 2005), which will need several evaluations and cost a lot. Usability tests are costly, but international usability tests cost even more (Law & Hvanneberg, 2004). For example, a Chinese company hopes to extend their products into Denmark. It may be less costly to use their own usability professionals to do the test in Denmark instead of employing Danish usability professionals. Considering the cross-cultural cost-benefit analyses (Lidwell, Holden, & Butler, 2003; Mayhew & Bias, 2005), it is worthwhile investigating how foreign evaluators will conduct the usability tests in the target country.
3) It may be more effective to use foreign evaluators when it is hard to get local expert evaluators. If using local evaluators in the target country, foreign evaluators who are from the software developing company have to train local evaluators about the way to do the tests for this application. The foreign evaluators need to watch the tests in the observation

room, and sometimes need to communicate with the local facilitator. The training and communication during the tests between the foreign evaluator and local facilitator are not easy to do (Dray & Siegel, 2005). Time and effort are needed. Sometimes the foreign evaluators who observe the tests in the observation room need to write down the usability problems and make the report to the company. Thus, in some situations, it may be more convenient and reasonable to use foreign evaluators instead of local ones.

From the above discussion, the use of foreign evaluators occurs in the usability engineering area. Since usability practitioners are not uniformly and homogenously distributed across the world, evaluators with different cultural backgrounds may have some specific features in communicating with users and finding and rating usability problems. This research investigates the impact of the evaluators' and users' cultural backgrounds on the result of the usability testing-usability problems, and on the process of the usability testing- evaluators' and users' communications.

1.3.2 Motivation from the Theoretical Point of View

Cognition and communication differences in various cultures (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005) call for a serious questioning of whether our knowledge and assumptions of the thinking aloud method is valid outside of Europe and the US. Speaking and language are closely related to human thinking, which is evident in the Western intellectual tradition, but it is not the case in the East Asian cultural tradition (Kim, 2002). East Asians believe that states of silence and introspection are considered to be beneficial for high levels of thinking (Kim, 2002). Thus, asking East Asian users to speak their thoughts while doing a task may influence their task performance more than it does for Westerners.

Hence, it is worthwhile investigating how to carry out thinking aloud usability testing in East Asian countries and whether there is any difference between the tests in East Asia and the West. Accordingly, this research investigates how Western and East Asian evaluators carry out the tests with the target users by using Danish and Chinese participants. Nisbett's culture theory designates Denmark to be a Western country and China an East Asian country. Moreover, according to Hall (1989, 1990), Denmark is a low-context communication country and China a high-context communication country. Thus, Denmark and China represent the two distinct cultures that the culture theory describes.

According to the work conducted by Sanchez-Burks, Nisbett, & Ybarra (2000), Westerners tend to have a task-focused orientation, which means that they focus on the task, not the socio-emotional climate, whereas East Asians tend to have a socio-emotional relational orientation, which means people's efforts and attentions are focused on the interpersonal climate of the situation. Westerners and East Asians may have different responses to local and foreign evaluators. Therefore, the current research is conducted in both Denmark and China with both Danish and Chinese evaluators and users. The study examines whether the cultural differences described by researchers (Hall, 1989a; Hall & Hall, 1990; Kim, 2002; Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002a; Sanchez-Burks et al., 2000; Spencer-Oatey, 2000) can be seen in usability testing situations.

1.4 Research Objectives

This research is aimed at getting a deeper understanding of the thinking aloud usability testing in the context of the intra- and inter-cultural usability engineering. The primary theoretical objective is to understand culture's impact on thinking aloud usability testing by applying Nisbett and his colleagues' culture theory of cognitive difference and Hall's cultural effects on communication (Hall, 1989a, 1990; Hall & Hall, 1990) and to determine a more valid thinking aloud model (Boren & Ramey, 2000; Ericsson & Simon, 1993) for the usability testing in different cultural settings.

The primary empirical objective is to examine the impact of the evaluators' and users' cultural backgrounds on the result (usability problems) and the process (communications) of the usability testing through a systematic study in Denmark and China with both Danish and Chinese evaluators and users. From this research, we hope to give valuable advice to usability practitioners on running tests effectively in the global market.

1.5 Research Method

An experimental study was conducted in Denmark and China to investigate the research question. The experimental design involved four independent groups: Danish evaluator-Danish user pairs, Chinese evaluator-Danish user pairs, Chinese evaluator-Chinese user pairs and Danish evaluator-Chinese user pairs. The tests with Danish users conducted in Denmark and the tests with Chinese users conducted in China. Usability professionals were invited to attend this research as evaluators. Foreign evaluators' travel fees were paid by the PhD project. In the usability tests, the testing application was a Danish/Chinese wedding invitation prototype and the general task was to ask the users to make an invitation for their wedding. The whole testing, including the faces of the evaluators and users, the screen and keyboard activities, was video- and audio- recorded by the software called Morae (section 3.4.3). Empirical data were collected by using background questionnaires, usability problem lists (for the users), usability problem forms (for the evaluators), video recordings of the testing and interviews. The issues of external validity and internal validity are addressed in section 6.3.

1.6 Related Work

There are many studies on the impact of culture on system usability (Aryee et al., 1999; Beu, Honold, & Yuan, 2000; Bilal & Bachir, 2007; Bourges-Waldegg & Scrivener, 2000; Callahan, 2004; Callahan, 2005; Cinnirella & Green, 2007; Duncker, 2002; Galdo & Nielsen, 1996; Hall, De Jong, & Steehouder, 2004; Sun, 2003; Sun, 2006; Zahedi, Van Pelt, & Srite, 2006). More specifically, Sun (2006) investigated cultural usability by comparing user localization efforts of mobile messaging technology in US and China. Honold (2000) identified 8 factors which may influence the use of products in foreign cultures and suggested considering those factors when designing products for different cultures. Bourges-Waldegg and Scrivener (2000) proposed Meaning in Mediated Action approach to design system/interface for culturally diverse use groups. Downey et al. (2005) investigated a possible relationship between the national culture and usability of an e-learning system by considering Hofstede's cultural dimensions and Nielsen's usability attributes. These studies show that cultural diversity of the target user groups should be taken into consideration in the system development. Since usability evaluation is an important stage in the system development and this research concentrates on culture's influence on UEMs, this study focuses on presenting the related work of the impact of culture on usability evaluation in section 1.6 in order to get a general view of this research area and to better understand the research topic.

Culture's influence on usability evaluation methods have been investigated by different researchers. Beu, Honold, & Yuan (2000, p. 355) put forward "different cultures, different evaluation methods," with the argument that there may be culture-related barriers if focus groups were to be held the same way in China as in Germany or the US. Chinese participants may be not willing to criticize others directly, but only indirectly. A common consensus, rather than individual differences, is supposed to be achieved. Guidelines of appropriate ways to lead focus group were provided in their study. For example, “Participants should talk about more than just their own experiences. Their working environment should also be included in the discussion (have you or your colleagues ever noticed any difficulties?)” and “Criticism of a product should be extolled as an opportunity for the manufacturer to learn” (Beu et al., 2000, p. 356).

Yeo (1998a) examined cultural factors that may affect the results of usability evaluation techniques. From his study, power distance was shown as an important cultural factor that influenced usability testing. The author found that a test user who was of higher rank than the experimenter gave more negative comments about the product than did the one who was of lower rank than the experimenter. In another study conducted by Yeo (2001a), he employed three usability assessment techniques: thinking-aloud technique (objective measure), system usability scale (SUS, subjective measure) (Brooke, 1996) and interviews (subjective measure). The results of the usability evaluations were found to be inconsistent. He found that for the less experienced computer users, or for the users who were not familiar with the evaluators, the objective measure and subjective measure were not matched. Even though these users performed poorly on the task, they still showed a positive attitude towards the software in the interview or SUS.

Lee and Lee (2007) explored cultural effects on the process and result of the usability evaluation techniques in Netherlands and Korea. They found that in the usability test, Dutch participants criticized the products more actively, and they discovered a product’s weakness and also its strength more frequently than did the Korean participants. Further, Dutch participants believed that most problems that occurred during the test were due to the problems with the product, whereas Korean participants believed that problems that occurred during the test were due to their mistakes. For the focus group interview, the results showed that Dutch participants actively engaged in a discussion soon after the interview started, whereas Korean participants took a while to start speaking up. In the Korean group, participants rarely spoke voluntarily before they were called upon by the moderator. The moderator needed to call on participants constantly and ask more detailed questions to carry on the discussion. On the other hand, the Dutch moderator did not have to do much because Dutch participants actively engaged in discussion, and some of them even had the tendency to speak too long which required the moderator to control such behavior.

The study conducted by Vatrapu and Pérez-Quiñones (2006) showed that even in structured interviews, when with foreign interviewers, Indian users were not willing to talk as freely and accurately as when with a local evaluator. Language may not be the key issue, since in their research both interviewers and users could speak English fluently. Their research found that the culture of the interviewer had an effect on the number of usability problems found, on the number of suggestions made, and on the number of positive and negative comments given. The local interviewer (Indian culture) brought more usability problems and made more suggestions than did the foreign interviewer (Anglo-American).

Yammiyavar, Clemmensen, & Kumar (2008) investigated the culture's influence on non-verbal communication in usability testing. They did 12 thinking aloud usability tests in Denmark, China and India with 4 tests in each country, and selected a total of 120 minutes of videos (10 minutes duration each) to analyze the non-verbal communication between the users and evaluators. The result showed that some non-verbal behaviors, such as "adapters" were significantly different in the three countries.

Clemmensen et al. (2007) did a cross-cultural field study of think aloud testing in seven companies in Denmark, China and India. They went to the companies that tested software and products for the local markets and observed how the usability professionals carried out the tests in different countries. They found that usability tests were not run in the same way in the three countries, as evident in: the attitudes toward users with different genders or ages were slightly different, the probing behaviors were different, the preferred ways of presenting the tasks were not the same, the experienced relations between the evaluators and users were different, etc. For the study in China, Shi (2008a) examined the relation and communication between the evaluators and users. She found that Chinese users did not think aloud actively, and thus the evaluators' effective communication skills were more important in the Chinese tests.

The above studies demonstrate that the way of carrying out UEMs does have specific features in different cultures. Accordingly, this research investigates how evaluators and users with the same and different cultural backgrounds communicate with each other to see the culture's impact on thinking aloud usability testing. Additionally, depending on the way of carrying out the UEMs, culture may also influence the result of the UEMs-identified usability problems.

In the study conducted by Law and Hvanneberg (2004), usability tests were performed by users from four European countries and local testers. Two evaluators extracted the usability problems by seeing the transcribed and translated thinking aloud protocols, the local testers' observation reports and the videos. The results showed that the heterogeneity of subgroups in a sample diluted the problem discovery rate, which meant that the usability problem discovery rate within all the uses from the four countries was much lower than the usability problem discovery rate within the users in one country. This study implicates that in order to find all the possible usability problems, compared to the usability tests for a specific culture, the usability testing involving users from different cultures may need to recruit more users.

Some studies also show that usability is not understood the same by users and software developers across cultures. Hertzum et al. (2007) did 48 repertory-grid interviews in Denmark, China and India to investigate the users' and developers' usability constructs. They found that for Chinese participants, security, task types, training and system issues were important in usability. For Danish and, to some extent, Indian participants, ease of use, intuitive and liked were important in usability. Frandsen-Thorlacius, Hornbæk, Hertzum, & Clemmensen (2009) conducted a questionnaire with a sample of 412 users from Denmark and China to investigate how they understood and prioritized different aspects of usability. The result showed that Chinese users prioritized visual appearance, satisfaction and fun higher than did the Danish users; however, Danish users tended to be more concerned with effectiveness, efficiency and lack of frustration than Chinese users. From their studies, we can see that culture influences the perception of usability, which implies that foreign evaluators may not find similar usability problems or may not rate the usability problem severity the same as local evaluators in the usability testing.

This section has reviewed the studies about the impact of culture on UEMs. Previous research indicates that culture does have influence on the way of conducting usability evaluation techniques and the identified usability problems. However, to this researcher's knowledge, there is no study investigating the cross-cultural thinking aloud usability testing systematically. This current research conducts the thinking aloud usability testing with both Danish and Chinese evaluators in both Denmark and China to investigate whether local and foreign evaluators determine different usability problems, how Danish and Chinese evaluators rate the severity of the problems, and how they communicate with the target users in order to find those problems.

From this research, we hope to develop a deeper understanding of the process and the result of the thinking aloud usability testing in cross-cultural settings.

1.7 Thesis Structure

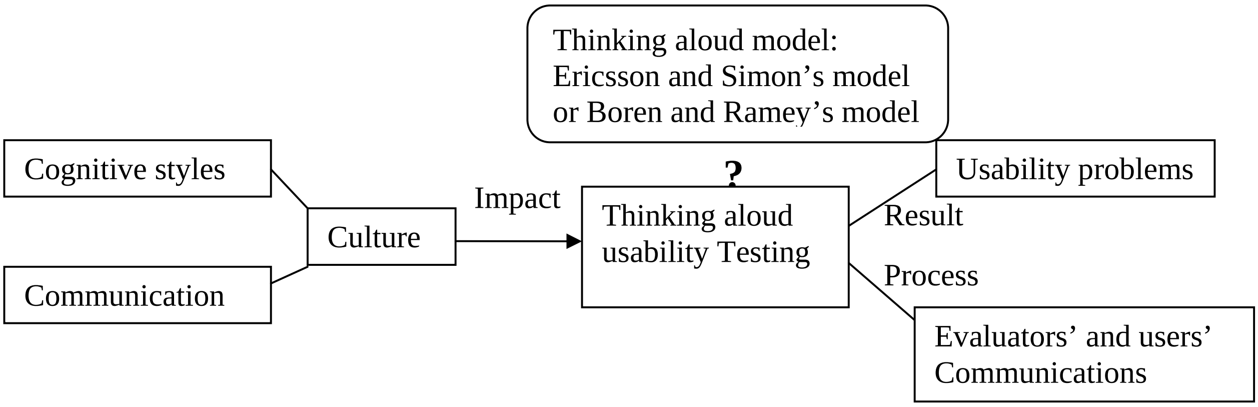
The dissertation is organized into six chapters. The outline of the rest of the thesis is as follows:

- Chapter 2 presents the theoretical background of this research, including the important concepts in the thinking aloud usability testing, thinking aloud theories and culture theories in this research. Moreover, culture's possible influences on usability testing are discussed when introducing the culture theories. Based on the discussion of the theories, hypotheses and themes for this research are put forward.
- Chapter 3, the methodology chapter, introduces the research design, participants, materials, procedures and data analysis.
- Chapter 4, the results, provides the demographic information of the participants, the statistic analysis of the findings and the interviews in this research.
- Chapter 5, the discussion, focuses mainly on a discussion of the findings in chapter 4.
- Chapter 6, the conclusion, includes a summary of the major findings of this research, the theoretical, methodological and practical implications, a discussion of the limitations, and recommendations for further research.

The introduction chapter gives a general view of this dissertation. The next chapter presents the theoretical background of this study.

2 Theoretical Background

This research investigates the thinking aloud usability testing in the intra- and inter- cultural settings, which involves evaluators and users with similar or different cultural backgrounds. From this dissertation, we hope to understand the extent to which evaluators' and users' cultural backgrounds impact the thinking aloud usability testing. This research holds an empirical realism perspective on usability testing (Bryman, 2008, p. 14). The world is able to be understood through using appropriate methods. Figure 2 presents the theoretical framework for this thesis.



Thinking aloud model: Ericsson and Simon's model or Boren and Ramey's model
Cognitive styles
Communication
Culture
Impact
?
Thinking aloud usability Testing
Result
Process
Usability problems
Evaluators' and users' Communications

Figure 2: Theoretical Framework

In this theoretical framework, we can see that culture may impact the thinking aloud usability testing. The thinking aloud method has been widely used in usability evaluation (Hertzum et al., 2009; Krahmer & Ummelen, 2004a), but the theories for thinking aloud in usability engineering are debatable (Boren & Ramey, 2000; Ericsson & Simon, 1993). In this chapter, the two main thinking aloud models are discussed: Ericsson and Simon's model (1993) and Boren and Ramey's model (2000). From this research, we hope to determine which model is more appropriate for thinking aloud usability testing.

Culture in this research is defined as different cognitive styles and communication orientations. Nisbett and his colleagues' findings of different cognitive styles across cultures (Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002b; Nisbett et al., 2001) and Hall's different communication orientations (Hall, 1989a, 1989b, 1990; Hall & Hall, 1990) are the culture theories followed in this dissertation.

Culture may impact many aspects of the thinking aloud usability testing, such as the way to prepare the tests, the way to design the tasks, etc. However, in this research, I mainly investigate the culture's impact on usability problems, and the evaluators' and users' communications, as described in the two sub-research questions.

This theoretical chapter is organized as follows:

- Section 2.1 Thinking Aloud Usability Testing: Introduces usability, usability testing, and the important concepts in the thinking aloud usability testing. Moreover, Ericsson and Simon's thinking aloud model and Boren and Ramey's thinking aloud model are discussed.
- Section 2.2 Culture and Usability: Illustrates the culture theories in this research and discusses the potential influence of culture on usability testing from the theoretical point of view.
- Section 2.3 Research Question, Hypotheses and Themes: Discusses the research question and sub-research questions, after which the hypotheses and themes are developed based on the theories.

2.1 Thinking Aloud Usability Testing

2.1.1 Usability and Usability Testing

2.1.1.1 Usability

Usability is one of the important quality characteristics of software systems and products (Jokela, 2004). It has been regarded as a science (Gillan & Bias, 2001; Lindgaard, 2009). There are many topics in the usability science, such as usability definition, usability design and usability evaluation (Benbunan-Fich, 2001; Frandsen-Thorlacius et al., 2009; Gray & Salzman, 1998b; Hartson et al., 2001; Hertzum et al., 2007; Hornbaek, 2006; Hornbæk, 2009; Hornbæk & Frøkjær, 2005). This thesis investigates one small area in the usability research—usability testing. Before introducing usability testing, we first need to know the concept of usability.

Usability, to some extent, is the question of "whether the system is good enough to satisfy all the needs and requirements of the users and other potential stakeholders, such as the users' clients and managers" (Nielsen, 1993, p. 24). Usability is defined 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 9241-11). Nielsen identifies the following components as the usability attributes (Nielsen, 1993, p. 26):

- Learnability: The system should be easy to learn so that the user can rapidly start getting some work done with the system.
- Efficiency: The system should be efficient to use, so that once the user has learned the system, a high level of productivity is possible.
- Memorability: The system should be easy to remember so that the casual user is able to return after some period of not having used it, without having to relearn everything.
- Errors: The system should have a low error rate so that users make few errors during the use of the system, and so that if they do make errors, they can easily recover from them. Further, catastrophic errors must not occur.
- Satisfaction: The system should be pleasant to use so that users are subjectively satisfied when using it; they like it.

From the definition and attributes of usability, we see that usability cannot be talked about without mentioning the users. The concept of usability is not limited to the products' attributes (Frandsen-Thorlacius et al., 2009). The users' use of experience and feelings of the system/product plays an important role in the usability definition and attributes. In this research, the definition of usability includes both the products' attributes and the users' experience. In order to learn from the users about the product's usability, usability testing will be conducted (Barnum, 2002).

2.1.1.2 Usability Testing

Usability testing is the most fundamental usability evaluation technique (Nielsen, 1993) and has now been accepted as an essential activity in the lifecycle of software design, implementation, testing, acceptance, and revision (Norman & Panizzi, 2006). It has become the primary UEM since the 1980s (Hartson et al., 2001). Through usability testing, the designers and usability engineers get direct information both about how people use the artifact and the problems that need to be improved in the next engineering lifecycle. Usability testing is viewed as an "empirical" method which means that the information comes from the real users (Barnum, 2002). Dumas and Redish summarize the characteristics of usability testing as (Barnum, 2002, p. 9):

- The primary goal is to improve the usability of a product. For each test, there must be specific goals and concerns that you articulate when planning the test.

- The participants represent real users.
- The participants do real tasks.
- The team observes and records what participants do and say.
- The team analyzes the data, diagnoses the problems, and recommends changes to fix these problems.

Usability testing is a user-involved method which means that it involves the users' task performing behaviors. Usability testing includes performance measurement and thinking aloud (Nielsen, 1993). Performance measurement is regarded as summative evaluation, whereas thinking aloud is regarded as formative evaluation (Nielsen, 1993, p. 170). Systems or products are usually designed through iterative process, which includes design, evaluation and redesign (Hartson et al., 2001; Mayhew & Bias, 2005). Before the final design is accepted for release, formative evaluation is often conducted in order to find usability problems of the prototype. By solving the problems in the next design stage, the usability of the prototype will be improved (Hartson et al., 2001). In contrast, summative evaluation is an evaluation of the final user interface, that is, to measure how well a product meets its stated usability goals, and how well it relies on quantitative metrics of effectiveness, efficiency, and satisfaction (Barnum, 2002; Capra, 2006; Leventhal & Barnes, 2007). According to Hartson et al. (2001, p. 375), usability evaluation methods are used "primarily for formative evaluation during the prototype design stage." Table 1 shows the testing subjects, method and purposes of the formative and summative evaluation by summarizing the literature (Barnum, 2002; Capra, 2006; Hartson et al., 2001; Nielsen, 1993).

Table 1: Formative and summative usability testing

Evaluation types	Testing subject	Method	Purpose
Formative evaluation	Prototype	Thinking aloud usability testing	Find usability problems in order to improve the prototype
Summative evaluation	Final product	Performance measurement	Assess the overall quality of the product

From Table 1, we can see that formative evaluation is done in order to help improve the product, and its goal is to identify a set of usability problems (Capra, 2006; Nielsen, 1993). Summative evaluation is done at or near the completion of the development cycle, and the purpose is to assess the overall quality of the product (Nielsen, 1993, p. 170), such as "measure how well the product performs against stated goals," or "help developers understand how users learn the new product" (Barnum, 2002, p. 122). Although summative evaluation is useful, a growing number of companies realize that it is necessary to get feedback from users through the whole application development cycle. Nearly all successful products today use user-centered design (Barnum, 2002). Hence, formative evaluation which involves users throughout the product development process is more and more important today. "Formative evaluation can begin even before there is a product to test, and it can continue late into the development process before the product is released" (Barnum, 2002, p. 123). The thinking aloud usability testing is a typical method used for formative evaluation (Nielsen, 1993, p. 170); it will be introduced in the following section.

2.1.2 Thinking Aloud Test

Over the last three decades, the thinking aloud method has become a popular instrument for studying cognitive processes, such as problem solving, human-computer interaction, reading and writing (Krahmer & Ummelen, 2004a). The thinking aloud usability testing involves having participants verbalizing their thoughts while performing a task using the system (Nielsen, 1993, p. 195). The basic principle of this method is that potential users are asked to complete a set of tasks with the tested product/ interface, and to constantly verbalize their thoughts while working on the tasks (Haak, Jong, & Schellens, 2003). In a thinking aloud usability testing, there are some important concepts or elements (Clemmensen et al., 2009): users, evaluators, tasks, the application/prototype being tested and usability problems. The relations of the important concepts in this research are developed in Figure 3.

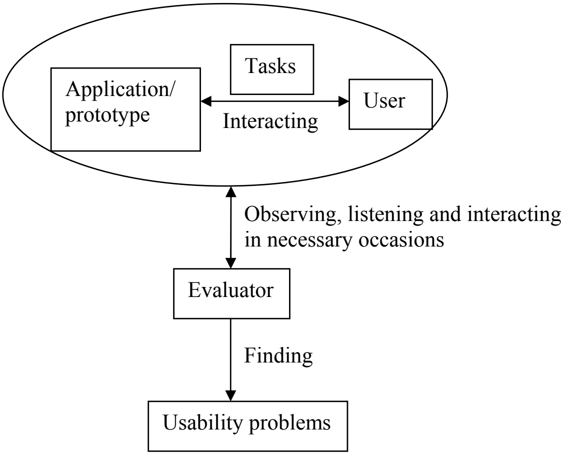


Figure 3: A relation model of the important concepts in a think aloud usability test

Figure 3 shows the important concepts of a think aloud usability test and the relations of the concepts. In this section, the important concepts are introduced and the relations between the concepts are discussed.

'User' is the focus of usability, and usability testing is the process of learning from users about a product's usability (Barnum, 2002). As Nielsen states, "individual user characteristics and variability in tasks are the two factors with the largest impact on usability" (Nielsen, 1993, p. 73). Users should be studied carefully, and recruiting the right users is one of the most crucial issues for the test (Kuniavsky, 2003, p. 265). When evaluating the usability of a product or interface in thinking aloud usability testing, it is necessary to recruit the target users to attend, since usability refers to the product's ability to fulfill target users' goals and needs with effectiveness and efficiency in a specified context of use (ISO 9241-11). The target users are the class of people who will be using the system (Nielsen, 1993, p. 74).

Evaluators are those who evaluate the systems/ products/ interface with a usability evaluation method and detect usability problems. In thinking aloud usability testing which involves end-users, evaluator's tasks are to give instructions and tasks to the user, facilitate the test, observe the user's task performance, interact with users and detect usability problems.

Task is regarded as one of the two most important issues (task and user) for usability (Nielsen, 1993, p. 43). Thinking aloud usability testing is generally a task-based session in a usability laboratory (Capra, 2006), where the user must do some tasks with the interface in the test (Norman & Panizzi, 2006). Users' tasks are not necessarily users' goals (Barnum, 2002, p. 91), as users' goals are things that users have. Through tasks, such as steps that the users take and processes that the users complete, the users' goals or objectives will be achieved. Users are primarily interested in achieving their goals, and the task is the means to it (Barnum, 2002, p. 92).

Application (or Product or Prototype) is the one being tested in usability testing. It is the subject of the test. The purpose of usability testing is to examine whether the application is good and what the problems are that may influence users' acceptance of the application and thus need to be modified. As introduced in section 2.1.1.2, usability evaluation includes summative and formative evaluations, so the test subjects in the two evaluations also have two kinds of applications: final product/application versus prototype (or unfinished product/application). The final product/application is evaluated by the summative evaluation method, whereas the prototype (or unfinished product/ application) is evaluated by the formative evaluation method. Further, there are two ways to develop an application for the international market: globalization and localization. Globalization seeks to "make products general enough to work everywhere" (Horton, 2005, p. 158), concentrating on separating the 'cultural elements' of a product from the rest of it (Horton, 2005); whereas localization seeks to "create custom versions for each locale" (Horton, 2005, p. 158), which is about adapting the cultural elements for a specific target culture (Horton, 2005).

Discovering the usability problems in the user interface is one of the most important goals and purposes of usability testing (Hartson et al., 2001; Law & Hvanneberg, 2004). Usability problem is defined by different researchers as:

- The parts of a system that cause users trouble, slow them down, or fit badly with their preferred ways of working, which are identified by systems developers or usability specialists through usability evaluation approaches (Hertzum & Jacobsen, 2001, p. 422).
- A problem experienced by the user, which is caused by an interaction flaw (Capra, 2006, p. 41).
- Any aspect of a user interface that is expected or observed to cause users problems with respect to some salient usability measure (e.g. learnability, performance, error rate,

subjective satisfaction) and that can be attributed to a single design aspect (Hvannberg et al., 2007, pp. 230-231).

- In software engineering, the problem involves some kind of system or process that requires a computerized solution. In usability engineering, the problem is to understand the user's tasks that must be supported by the user interface, in the context of the user and task characteristics. In other words, the variables that influence usability for a given project need to be identified (Leventhal & Barnes, 2007, p. 57).
- A usability problem is real if it predicts a problem that users will experience in their own environment, which affects their progress toward goals and their satisfaction (Wilson, 2007, p. 46).

The usability problem is related to the definition of usability. For different evaluators, their understanding of an application's usability may be different. Some evaluators may consider operation to be important, such as whether it is easy to use or whether it is easy to make errors; while others may consider satisfaction to be important, such as whether they like the interface color or not. Thus, the usability problems in this study should be clarified to every evaluator before the test. As discussed in section 2.1.1.1, usability in this study includes both the products' attributes and the users' experience, so usability problems are those factors that cause the system to be hard to operate or not to be satisfied by the users.

Figure 3 shows the relations of the important concepts in a thinking aloud usability testing. From figure 3, we can see that the way to evaluate an application or prototype is to ask users to conduct some tasks to see whether they can complete the tasks effectively and efficiently with the application, and whether they are satisfied with the application. The task is the means to evaluate the application. When doing tasks in the test, if the concept metaphor of the interface/product does not match the users' mental model, the users will make mistakes frequently. If the task flow of the interface/product matches the users' mental model, the users will be satisfied with it. The evaluator observes the user's interaction with the application and task performing. From observing the whole situation and listening to the users' "think aloud," the evaluator tries to find the problems of the application.

Advantages and disadvantages of the thinking aloud usability testing

Thinking aloud usability testing has both advantages and disadvantages. The strengths of the thinking aloud usability test are:

- Thinking aloud usability testing shows what the user is doing and why he/she is doing it while he/ she is doing it in order to avoid later rationalizations (Nielsen, 1993, p. 196).
- From the thinking aloud procedure, the focus of attention can be acquired when the users verbalize what they are reading or looking at, which is important since it involves the locus of information acquisition or the point of decision-making (Norman & Panizzi, 2006, p. 248).
- Thinking aloud tests can collect a wealth of qualitative data from a fairly small number of users.
- The users' comments often contain vivid and explicit quotes that can be used to make the test report more readable and memorable (Nielsen, 1993, p. 195).
- The verbal protocols that are elicited in the tests can be very rich in diagnostic and evaluative information (Norman & Panizzi, 2006, p. 248).

The main disadvantage of the thinking aloud usability testing is that it does not lend itself very well to most types of performance measurement (Nielsen, 1993, p. 195), which means it may influence the users' performance when they are speaking out aloud. It may also change the course of the user's behavior because of the demand of verbalization while performing (Norman & Panizzi, 2006, p. 248). Further, evaluators may not always be able to rely on the user's statement; instead, they need to make notes of what the users are doing. "Data showing where users actually looked has much higher validity than users' claim that they would have seen the field if it had been somewhere else" (Nielsen, 1993, p. 196). Third, the verbal protocols can be contrived, biased and misleading (Norman & Panizzi, 2006, p. 248).

The discussion of the advantages and disadvantages of the thinking aloud usability testing indicates that the testing is good for getting direct information from the users about the application/ prototype, but the thinking aloud data may not always be valid. In the next section, the thinking aloud models which illustrate the theoretical issues of the thinking aloud methods are introduced.

2.1.3 Thinking Aloud Models

The thinking aloud method has been increasingly used for studying cognitive processes in many areas of psychology, education and cognitive science since the 1980s. In a thinking aloud study, participants are asked to carry out a task while verbalizing their thoughts, and it is assumed that the verbal data can be trusted. Ericsson and Simon (1993, p. xii) claim that the verbal data is valid since "individuals had privileged access to their experiences." Thinking has been viewed as a temporal sequence of mental events (Ericsson & Simon, 1993), so thought processes can be described as "a sequence of states, each state containing the end products of cognitive processes, such as information retrieved from long-term memory, information perceived and recognized" (Ericsson & Simon, 1993, p. xiii). The information in a state is relatively stable, so it can be verbalized and reported orally (Ericsson & Simon, 1993, p. xii). In this section, the theoretical basis for Ericsson and Simon's thinking aloud method and Boren and Ramey's revised thinking aloud model for the usability testing field are introduced.

2.1.3.1 Ericsson and Simon's Classical Thinking Aloud Model

The classical thinking aloud theory was developed by Ericsson and Simon (1993), based on the information processing theory. The authors propose that the thoughts that are verbalized are simply the information that participants attend to while generating the answer. However, in the classical thinking aloud model, valid verbalization data does not include describing or explaining what they are doing. When the participants are asked to verbalize only their thoughts entering their attention while they are doing the task, the thought sequences and thought structures will not be changed (Ericsson, 1998; Ericsson & Simon, 1993). However, if the participants are asked to describe or explain their thoughts, such as stating the reasons and motives of the action, the sequence and structure of the thoughts will be changed because the additional information and thoughts are not needed to perform the task (Ericsson, 1998; Ericsson & Simon, 1993; Hayes, 1986).

Ericsson and Simon have presented a specific processing model to interpret verbal data. An important and more specific assumption is that information is stored in several memories which have different capacities and accessing characteristics. Within the framework of the information processing model, it is assumed that information recently acquired by the central processor is kept in short-term memory (STM), and is directly accessible for further processing (such as for producing verbal reports), whereas information from long-term memory (LTM) must first be retrieved or transferred to STM before it can be reported (Ericsson & Simon, 1993, p. 11). "Any verbalization or verbal report of the cognitive processes would have to be based on a subset of the information held in STM or LTM." (Ericsson & Simon, 1993, p. 12) The information is reliable if it is in STM in the concurrent thinking aloud test.

Ericsson and Simon describe three different levels of decreasingly reliable verbalization, where each level is characterized by the amount of interference caused by non-task related processing (Ericsson & Simon, 1993, p. 79):

- The first level of verbalization is that which needs not to be transformed before being verbalized during task performance. When information is reproduced in the same form as it was attended, it is level 1 verbalization. At this level, there are no intermediate processes, and the participant does not need to expend special effort to communicate his/her thoughts, such as verbalizing numbers, since it is in the same form as they were originally encoded in STM.
- The second level of verbalization is that which needs to be recoded before being verbalized, such as verbalizing images or abstract concepts, but it does not bring new information into the focus of the participant's attention. "The recoding does not change the structure of the process for performing the main task." (Ericsson & Simon, 1993, p. 79)
- The third level of verbalization is that which requires additional cognitive processes to explain the thought process or thoughts, such as the explanation of thoughts, ideas, or hypotheses which require the participant to make intermediate inferences and change the normal flow of information in the STM. The information in level 3 is not attended by the participant at first, and does not belong to the normal course of task performance. Examples of additional cognitive processes include "making inferences about the subjects' own cognition, information retrieved from long-term memory at the researcher's request, and any outside influence, such as any comment or prompt from the researcher." (Boren & Ramey, 2000, p. 262)

To summarize, with level 1 and level 2 verbalization, no additional information is added, and the sequence of the heeded information is not changed. On the other hand, level 3 verbalization needs additional information, which changes the sequence of the information that the participants attend.

Doing thinking aloud tests according to Ericsson and Simon's thinking aloud model

Performing thinking aloud tests according to Ericsson and Simon's thinking aloud model complies with the following rules (Boren & Ramey, 2000; Ericsson & Simon, 1993):

- In order to make the participants familiar with the thinking aloud method and the lab environment, it is better to have a warm-up procedure to train the participants to conform the thinking aloud instructions.
- The social interaction between the participant who is doing thinking aloud and the experimenter should be minimized. The experimenter should use "keep talking" instead of a social request such as "tell me what are you thinking" to remind the participant to think aloud when he/she keeps silent for a longer period of time. The intervals of silence are generally standardized, for example, after 15 seconds to 1 minute pauses, they are given a reminder.
- Except for the reminders, the experimenter should not interfere during the thinking aloud process.

The main use of thinking aloud method

Krahmer and Ummelen (2004b, pp. 1-2) summarize the main use of thinking aloud method to three types of goals:

- To find evidence for models and theories of cognitive processes;
- To discover and understand general patterns of behavior in the interaction with documents or applications, in order to create a scientific basis for designing them;
- To test specific new documents or applications in order to trouble-shoot and revise, such as doing usability testing, or pretesting, or formative testing. The primary goal is to gather user information to support the design of a specific product.

Ericsson and Simon's research matches the first category. They developed a theoretical framework and accordingly, a procedure for collecting valid and reliable thinking aloud data. The third goal is usually used in usability testing. However, the thinking aloud data which is considered to be "valid" in the classical model does not fulfill the third goal (Boren & Ramey, 2000). Usability practitioners do not follow the classical model to do the thinking aloud usability testing (Boren & Ramey, 2000; Nørgaard & Hornbæk, 2006; Shi, 2008a). Considering the discrepancies between the observed thinking aloud testing and Ericsson and Simon's theoretical basis, Boren and Ramey (2000) proposed a new approach based on speech communication theory.

2.1.3.2 Boren and Ramey's Thinking Aloud Model

Boren and Ramey (2000) conducted seven primary observations at two software companies and found that the strict guidelines prescribed by Ericsson and Simon (1993) were hardly complied with in practice. Therefore, a 'speech communication' paradigm that allows evaluators more freedom to interact with the users is proposed. This is motivated by a review of the differences in purpose between research into cognitive processes and research into usability testing. Boren and Ramey give the reasons why thinking aloud used in investigating cognition is different from that which is used in usability testing (Boren & Ramey, 2000, p. 263):

- First, the main concern of using thinking aloud method in cognitive psychology and usability research is different. In cognitive psychology, thinking aloud is mainly used to understand "how cognitive processes function in specific domains" (Boren & Ramey, 2000, p. 263). In the usability area, the primary concern is to "support the development of usable systems by identifying system deficiencies; building robust models of human cognition is not a central concern" (Boren & Ramey, 2000, p. 263).
- Second, there are different ways of obtaining important cognitive processes in cognitive psychology and usability research. In cognitive psychology, cognition is manifest through verbalization, whereas in usability testing, cognition that is of interest is shown not only through verbalization, but also through interactions with the application.
- Third, the application being tested in usability testing and the test apparatus used in a cognitive study are different. In usability testing, the application is more complex and may "contain many experimental unknowns that are difficult to control" (Boren & Ramey, 2000, p. 263).

The differences in research goals and contexts between cognitive psychology and usability testing imply that "the generalizability of Ericsson and Simon's theory to human factors and usability research was long overdue" (Boren & Ramey, 2000, p. 263). Ericsson and Simon's theory focuses primarily on cognitive processes, like problem solving. However, in the usability test, the main purpose is not only to get the user's cognition, but more importantly, to know the application and get the users' using experience and feelings of the application in order to improve the usability of a product. Thus, as long as the evaluator does not force his/her own opinion on the user, it is appropriate to build a supportive relationship and communication with the users in usability testing.

Boren and Ramey's approach focuses on the communication between the evaluator and test user. In the practice of usability testing, there is always a user and an evaluator. "Talk is not simply a form of action" performed by the user alone, "but a mode of interaction" between users and evaluators (Boren & Ramey, 2000, p. 267). "An utterance is not self-sufficient but is a link in a genre-based communicative chain" (Smagorinsky, 1998, p. 168). The users do not ignore the evaluator. They expect a response, agreement, sympathy, etc., from the evaluator. In the speech communication model, some key issues should be clarified:

- The subject of the test is the interface, not the user.
- The test user is the expert, who is assumed to provide valuable information of the interface. The evaluator is the learner, whose main task is to get information from the user's speech and find usability problems.
- The evaluator should use undirected and undisturbed tokens to make the user stay focused on the tasks and thus verbalize thoughts fluently.
- When encountering contingencies during the usability test, interaction between the evaluator and test user is required.

Having good communication is thus very necessary for usability testing (Shi & Clemmensen, 2007). Clemmensen and Shi (2008; 2007) and Shi (2008a) carried out some field studies in Denmark, India and China, observing the usability practitioners during the usability test in the industrial area. In all the three countries evaluators did not listen passively, but actively interacted with the users on necessary occasions to get more valuable information about the interfaces. So compared to Ericsson and Simon's classic thinking aloud model, Boren and Ramey's thinking aloud model may be more suitable for usability tests.

2.1.3.3 Discussion of Ericsson and Simon’s Thinking Aloud Model and Boren and Ramey’s Thinking Aloud Model

The two thinking aloud models describe different ways in doing thinking aloud usability testing. When following different models, usability practitioners do the usability tests differently, and the two models may fit for different kinds of usability evaluations. However, the two models also have a common feature, namely, that neither of them have much discussion of the “cultural issues.” Smagorinsky (1998, 2003) suggests that the thinking aloud protocol should be reconsidered from a cultural perspective. In this section, the applicability of the two models and the ignored cultural issues are discussed.

Applicability of the two models for thinking aloud usability testing

Ericsson and Simon’s model and Boren and Ramey’s model may fit for different types of thinking aloud usability evaluations. As discussed in 2.1.1.2, there are two kinds of evaluations: formative evaluation and summative evaluation. Although the thinking aloud usability testing is often used for formative evaluation, and a measurement test used for summative evaluation (Nielsen, 1993, p. 170), sometimes thinking aloud can also be used for summative evaluation to get the user’s cognition and thoughts of doing the task. But in the summative evaluation, in order not to influence the users’ performance, the evaluator should not interrupt the user during the test. According to Ericsson and Simon’s thinking aloud model (1993), the evaluator cannot interact with the user and he/she can only use “keep talking” to remind the user to talk. So the classic thinking aloud model described by Ericsson and Simon (1993) may be more suitable for summative usability tests (Shi, 2008a).

On the other hand, Boren and Ramey’s (2000) model may be more appropriate for formative evaluations. The main purpose of formative evaluation is not to see the overall performance of doing the task using the product/ interface, but to find the problems of the product/ interface and fix them in the next design cycle. In order to get the reliable usability problems, evaluators need to ensure that they got the users’ real feelings and thoughts, indicating that an effective communication with the users is necessary (Boren & Ramey, 2000; Tamler, 2001). Since the thinking aloud is often used as formative evaluation (Nielsen, 1993), Boren and Ramey’s (2000) model seems to be more important than Ericsson and Simon’s in the think aloud usability tests.

Therefore, whether evaluators should interrupt the user depends on the goals of the test (Dumas & Redish, 1999, p. 295). If it is important for the user to work alone in the test, such as recording the time and steps on the tasks, or to find problems that would cause users to call the company's customer support line (Dumas & Redish, 1999, p. 295), interruptions should be fewer. If testing an early prototype, the designers usually want to get "as much diagnostic information as possible" (Dumas & Redish, 1999, p. 296). In this situation, evaluators will have more interactions with the users. For example, evaluators may ask users to try another way to do the task, or ask them to explain why they did the task in the way they did. Evaluators often prepare a set of probing questions to ask users when they do not give comment that they want to get or when they do not notice some important parts that the designers are interested in with early prototypes in formative evaluations. In this research, a prototype, rather than a finished application, is the testing subject; common use of the thinking aloud usability testing-formative evaluation, rather than summative evaluation, is conducted, and thus Boren and Ramey's thinking aloud model is the main concern.

Considering the thinking aloud models in usability testing from a cultural perspective

Cultural issues are not the main concern in Ericsson and Simon's, as well as Boren and Ramey's model. However, culture is an important issue for the cultural psychological tradition. Smagorinsky (2003, p 234) argues that the study of cognition "cannot be isolated from its social and cultural relationships." In the two models, we can see examples of socially directed speech. In Ericsson and Simon's thinking aloud model, the level 3 verbalization which includes the explanation of the thought or thought process could be regarded as involving socially directed speech (Ericsson, 1998). Boren and Ramey's thinking aloud model, which is developed based on the speech-communication theory, also implicates a link to other people. However, the two models do not explicitly discuss the cultural impact on thinking aloud.

Smagorinsky (2003) considered the cognitive process as a consequence of cultural practices which varied in different cultural groups. He discussed the protocol analysis from a cultural perspective (Smagorinsky, 2003). Usability testing is regarded as a cognitive activity (Hertzum & Jacobsen, 2001), thus Smagorinsky's (2003) theoretical discussion of the thinking aloud usability testing from a cultural perspective becomes of interest. Figure 4 shows the relation model of the important concepts in the thinking aloud usability testing by considering cultural issues.

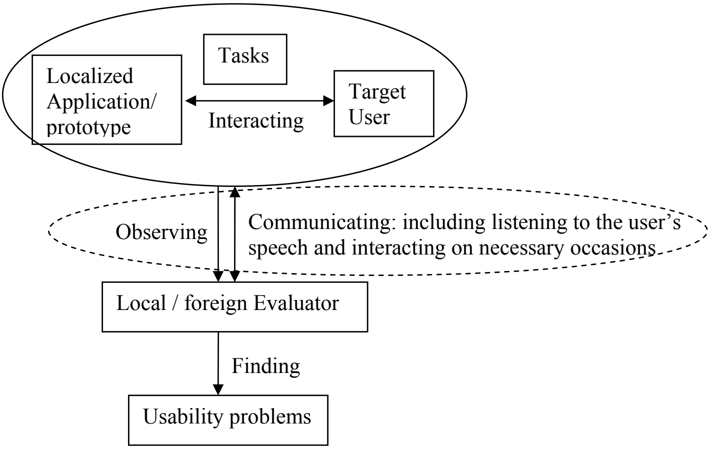

Tasks
Localized Application/ prototype
Interacting
Target User
Observing
Communicating: including listening to the user's speech and interacting on necessary occasions
Local / foreign Evaluator
Finding
Usability problems

Figure 4: A relation model of the important concepts in a thinking aloud usability test by considering cultural issues

Figure 4 is developed on the basis of Figure 3. The cultural issues include the evaluators' cultural background, localized application/ prototype and target users. If the users, tasks and applications are similar when doing tests in one country, and the only difference is using local or foreign evaluators, from Figure 4 we can see that cultural issues may influence the evaluators' observation and communication. Evaluators with different cultural backgrounds may have different observations for the users' task performance and communication, which may result in finding different usability problems or rating the severity of the usability problems differently.

Ericsson and Simon's model and Boren and Ramey's model do not have much discussion about the cultural issues. From this PhD research, we hope to get a better understanding of thinking aloud usability testing in different cultural settings.

2.1.4 Summary

In section 2.1, usability, usability testing and thinking aloud usability testing have been introduced. Usability in this research includes both the products' attributes and the users' experience. The important concepts in a thinking aloud usability testing, i.e. users, evaluators, tasks, application and usability problems have also been briefly introduced. Regarding the discussion of Ericsson and Simon's thinking aloud model, as well as Boren and Ramey's thinking aloud model, there are not only the applicability issues, but also the ignored cultural issues. However, in order to understand the "culture issues," what "culture" is needs to be clarified. In the next section, the definition of culture, the culture theories in this research, and the culture's potential influence on the thinking aloud usability testing are presented.

2.2 Culture and Usability

This PhD thesis is a cross-cultural research, and thus we need to establish a common understanding of the concept of culture and how it is understood in the human-computer interaction (HCI) area. In this section, the definition of culture and the culture theories used in this study are discussed.

2.2.1 Definition of Culture in this research

Culture has been defined in many different ways by different researchers. The definitions of culture which have been used in cultural usability are introduced here. Yeo (1996) thought culture could be defined by country boundaries, language, cultural conventions, race, shared activities or workplace. Smith and Yetin (2004, p. 1) states that "when defining culture, researchers often refer to patterns of values, attitudes, and behaviors, which are shared by two or more people." According to Geert Hofstede, culture is defined as "the collective programming of the mind that distinguishes the members of one group or category of people from another" (Hofstede, 2001, p. 9). Hofstede's culture theory is used by many researchers in the HCI area (Callahan, 2005; Marcus, 2005; Marcus & Gould, 2000; Vatrapu & Pérez-Quiñones, 2006; Vatrapu, 2007; Vöhringer-Kuhnt, 2002; Zahedi et al., 2006). However, this theory is more value orientated (Hofstede, 2001) and the above definitions are given more from the sociological point of view. In usability testing, a more psychological and more process orientated definition is needed. Since usability testing is a cognitive activity (Hertzum & Jacobsen, 2001), Nisbett's view of culture, which is about people's cognition and perception difference, is used as one of the main culture theories in this study.

According to Nisbett and his colleagues (Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett et al., 2001), people in different cultures perceive the world differently. Nisbett and Norenzayan (2002a, p. 3) claim that "cultural practices and cognitive processes constitute one another. Cultural practices encourage and sustain certain kinds of cognitive processes, which then perpetuate the cultural practices." In Nisbett's culture theory, Europeans and Americans (Westerners) are analytical thinkers, whereas Chinese, Korean and Japanese (East Asians) are holistic thinkers. Westerners and East Asians hold different cognitive styles. Cognitive style is the preferred way to perceive, think and remember information and use the information to solve problems. It is defined as "a characteristic and self-consistent mode of intellectual and perceptual functioning" (Colman, 2001). Based on the consideration of Nisbett and his colleagues' studies, culture is defined as cognitive style in this research.

Even though cognitive style is important in usability testing, communication also plays an important role in understanding users' opinions and feelings of the application. In this research, Hall's high-context and low-context communication orientations are applied as well. Cognitive styles and communication are not two separate concepts, but related to each other. A study shows that people with holistic and analytical cognitive styles prefer different communication strategies (Littlemore, 2001). Littlemore (2001) found that holistic students used more communication strategies that were based on comparison, and analytical students used more strategies that were based on individual features of the target item. According to Hall, "culture is communication" (Hall, 1990, p. 94). People with different cultural backgrounds may have problems in communicating with each other due to the interdependencies between languages, cognition and interaction (Gumperz & Levinson, 1991). Low contextual and high-contextual communications are put forward by Hall (1989a, 1990). People with low-contextual communication tend to describe in detail since the context information is supposedly unknown by the listener. On the other hand, for people with high-context communication, some information is not necessary to be conveyed since it is already known by the listener. In thinking aloud usability tests, evaluators and users with different cultural backgrounds may communicate differently in order to find relevant usability problems. So culture is also defined as communication in this research.

In a word, in this research, culture is cognitive style and communication. In the following sections, the two dimensions are discussed in greater detail.

2.2.2 Culture is Cognitive Style

Richard Nisbett and his colleagues proposed a powerful cognitive perspective on culture and behavior (Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Norenzayan, 2002b). In Nisbett's theory, there are many cognition and perception differences between Europeans, Americans, Chinese, Korean, and Japanese. Before introducing Nisbett's culture theory and discussing the relation between this theory and usability testing, cognitive style is briefly introduced.

2.2.2.1 Cognitive Style

Cognitive style is "the characteristic self-consistent modes of functioning found pervasively throughout an individual's cognitive, that is, perceptual and intellectual, activities" (Witkin, 1967, p. 234). It is also defined by Riding and Rayner (2000) as the person's typical, preferred way or manner in processing, organizing information, solving problems, and learning. Kozhevnikov (2007, p. 477) articulates that cognitive styles "represent heuristics an individual uses to process information about his or her environment."

People's cognitive styles are related to their family experiences while growing up (Witkin, 1967). Longitudinal studies indicate that cognitive style can be quite stable over three and a half years (Clapp, 1993, cited from Bull & McCalla, 2002). There are many kinds of cognitive styles, such as convergence-divergence, holistic-analytic, field dependence-independence, intolerance of ambiguity, assimilator-explorer, locus of control, reflection-impulsivity, etc (Bull & McCalla, 2002; Colman, 2001; Riding & Rayner, 2000). Cognitive style plays an important role in interacting with other people. Many previous researchers (Bull & McCalla, 2002; Hayes & Allinson, 1996; Mullany, Tan, & Gallupe, 2007; Riding & Rayner, 2000; Sadler-Smith & Riding, 1999) have investigated the effects of matching and mismatching people's cognitive style with learning performance/ learning experience. Hayes and Allinson (1996) reviewed 19 studies and found that 12 of them supported the hypothesis that matching cognitive style with learning activity would have a positive effect on the learning performance. People with matched cognitive styles will feel comfortable when learning knowledge and interacting with others.

In this research, cognitive styles are discussed from a cultural perspective, since they are connected to culture (Faiola & Matei, 2005; Gutierrez & Rogoff, 2003). Culture influences the contents of thought and shapes the cognitive styles (Cole, 1996; Nisbett & Norenzayan, 2002b). The influence of culture on cognitive styles can be seen from the linguistic relativity hypothesis (Gumperz & Levinson, 1991; Lucy, 1992a, 1992b; Nisbett & Norenzayan, 2002a). The linguistic relativity hypothesis is "the hypothesis that differences among language in the grammatical structuring of meaning influence habitual thoughts" (Lucy, 1992a, p. 1). Whorf argued that "as languages differ, so do the thoughts of the people who use them" (Gumperz & Levinson, 1991, p. 324). Both Danish and English belong to the Germanic branches of the Indo-European language family and the two languages are very close in many aspects, whereas Chinese belongs to Sino-Tibetan language family. Even though many studies have compared the American and Chinese people's behaviors from the linguistic relativity hypothesis, the findings of those studies might be able to give implications to the situation in Denmark and China to some extent. In order to better connect previous researchers' findings with this study, the words of "Western" and "East Asian" are used in this study.

The cultural cognitive style does play an increasing role in multinational software developing countries. Designers' cognitive style influences the usability. Faiola and Matei (2005) have explored the relation between web designers' cultural cognitive styles and the impact of the interface they designed on user responses. Websites created by both Chinese and American designers were given to American and Chinese users. The results showed that users performed information-seeking tasks faster when using web content created by designers from their own cultures. The study indicates that the designers' cultural cognitive styles influence the web design, and the designed websites also affect the target users' performance.

However, whether the cultural cognitive style also influences usability testing is not so clear now. Since cognitive styles influence the people's interaction (Bull & McCalla, 2002), it implies that cognitive styles may also influence the evaluator's and user's interaction in the usability tests. Users' information processing habits and verbalizing approach involve their typical mode of perceiving, thinking, problem solving and verbalizing. At the same time, evaluators also have their own perceiving, thinking and understanding of users' behavior and talk. The evaluator and user need to communicate everything clearly. The whole procedure looks simple, but it involves many cognitive processes. If the cognitive styles between the evaluator and user are matched, it should be easier for them to interact, or else, extra work might be needed.

Nisbett and his colleagues investigate cognitive style differences between Westerners and East Asians. In this study, the two main cognitive styles between East Asians and Westerners are discussed: holistic cognitive style versus analytic cognitive style.

Holistic Cognitive Style versus Analytic Cognitive Style

In Nisbett's theory, there are reliable differences in the modes of thinking between people from the East and the West. Generally, East Asian people's cognitive style is holistic and Western people's cognitive style is analytic (Nisbett, 2003; Nisbett & Miyamoto, 2005). Holistic- analytic cognitive style dimension refers to the extent to which a person processes information in wholes or separate parts (Bull & McCalla, 2002).

In Nisbett's culture theory (Nisbett et al., 2001, p. 293), holistic cognitive style is defined as "involving an orientation to the context or field as a whole, including attention to relationships between a focal object and the field, and a preference for explaining and predicting events on the basis of such relationships." Holistic approaches are dialectical, which means "there is an emphasis on change, a recognition of contradiction and of the need for multiple perspectives, and a search for the 'Middle Way' between opposing propositions" (Nisbett et al., 2001, p. 293). This approach relies on experience-based knowledge. People in the East Asian cultures of China, Korea, and Japan, have a relatively holistic cognitive orientation, seeing the world as interpenetrating and continuous, and emphasizing relationships and connectedness (Ji, Peng, & Nisbett, 2000).

Analytic cognitive style is defined as "involving detachment of the object from its context, a tendency to focus on attributes of the object to assign it to categories, and a preference for using rules about the categories to explain and predict the object's behavior. Inferences rest in part on the practice of decontextualizing structure from content, the use of formal logic, and avoidance of contradiction." (Nisbett et al., 2001, p. 293) People in Western countries, such as Europe and North America, have a relatively analytical cognitive orientation, seeing the world as being composed of discrete object or atoms and searching for the attributes of the object (Ji et al., 2000).

Originated from the different cognitive styles (Holistic and Analytic) in the West and East Asia, some concrete cognition differences have been found between Westerners and East Asians (Ji et al., 2000; Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002b; Nisbett et al., 2001). In the next sections, the cognition differences which have been investigated by Nisbett and his colleagues are discussed, with the consideration of their potential influence on the thinking aloud usability testing.

2.2.2.2 Nisbett's Culture Theory and Thinking Aloud Usability Testing

Nisbett and his colleagues' studies show different cognition orientations across cultures:

- Field independence vs. field dependence: East Asians are inclined to focus their attention broadly on the field, and find it more difficult to make a separation between an object and the field than do Westerners. (Nisbett & Masuda, 2003, p. 11165)
- Attention to object vs. field: East Asians attend to objects in relation to the field, and will be more thrown off by a presentation of the object against a novel field than would Westerners. Westerners are inclined to attend to focal objects. (Nisbett & Masuda, 2003, p. 11166)
- Causal Attribution, prediction and postdiction: East Asians are inclined to explain events by reference to interactions between the object and the field, while Westerners are inclined to explain events by reference to properties of the object (Nisbett & Masuda, 2003, p. 11164). With regard to postdiction (such as post hoc explanation), East Asians are not as surprised by unanticipated outcomes as Westerners (Nisbett et al., 2001, p. 299).
- Categorization: East Asians classify objects and events on the basis of relationships and family resemblance, whereas Westerners classify them on the basis of rule-based category membership. (Nisbett & Masuda, 2003, p. 11164)
- Perception of control: East Asians and Westerners respond differently to being given control in a situation. A sense of personal control is more important for Westerners than it is for East Asians (Ji et al., 2000, p. 944). When situational elements that normally signal control are present, Westerners believe they have more control, their performance improves more, and their confidence increases more than does that of East Asians.

- Logic vs. dialectics: East Asians are more comfortable with apparent contradictions than Westerners when presented with evidence for apparently contradictory propositions. When presented with apparent contradictions, Westerners resolve the situation by deciding which of the two propositions is correct, whereas East Asians are inclined to find some truth in both propositions. (Nisbett & Masuda, 2003, p. 11164)
- Task-focus and Socio-emotional Orientations: East Asians' effort and attention are directed toward the interpersonal climate of the situation, and they strive to maintain social harmony, while Westerners' effort is directed toward task-related goals, and attention is focused on monitoring the extent to which these goals are being accomplished. (Sanchez-Burks et al., 2000, p. 4)

In the following sub-sections, the cognition orientations are explained and the implications of the cognition differences on the thinking aloud usability testing are discussed.

Field Independence vs. Field Dependence

Field independence-dependence is considered to be the most widely known cognitive style according to Bonham (1987) (Heineman, retrieved in Jan. 2008). This dimension may influence the way to communicate with users in thinking aloud usability testing.

The Rod and Frame Test conducted by Ji et al. (2000) indicated that East Asians were inclined to have a field dependent cognitive style and Westerners were inclined to have a field independent cognitive style. The Rod and Frame Test (Witkin, 1967) is a classical test to examine the field dependence degree. The task is to report when the rod appears to be vertical. Ji et al.'s study (2000) showed that East Asians were more influenced by the frame position than were Westerners.

The dimension of field independence vs. dependence not only plays a role in separating an object from the field, but also influences the way people interact with others. Field independent people depend more on self and they appear to be more adept with the unstructured environment than are their field dependent counterparts. A field dependent person is reliant on external referents; dependent learners rely more on the teacher and peer support. The independent student tends to be more analytical and attends less to peer pressure or teacher direction (Heineman, retrieved in Jan. 2008).

In thinking aloud usability testing, how to communicate and support field dependent users may be more important than it is for field independent users, since field dependent users may rely more on evaluators and the testing environment. For the tests with East Asian users, the evaluators' skills of facilitating the test and communicating with users are inclined to be more highly required than those with Western users. Since Westerners tend to be field independent and may rely less on other people, the evaluators with Western users should always realize that users are the main speakers, and they should try to increase the users' roles and reduce their own roles.

Attention to the Object vs. Field

East Asians pay more attention to the field and Westerners pay more attention to objects when they are presented the same scene (Nisbett & Masuda, 2003). The attention differences may influence the identified usability problems in a thinking aloud usability testing. In this section, first, East Asians' and Westerners' attention differences are introduced, and then the influence of attention differences on usability testing is discussed.

A study conducted by Nisbett & Masuda (2003) found that the first statement by Japanese participants usually referred to the context, whereas the first statement by American participants usually referred to salient objects. In a subsequent recognition task, participants were shown a number of objects, some of which were in the original background and some of which had no background or were in a novel background. The participants were asked to identify whether or not they had seen the object before. The result showed that the Japanese made more errors when the object was seen against a novel background than when it was seen against the original background. There was no difference to the accuracy of the American participants. This indicates that the perception of objects by East Asians is usually connected with fields, whereas the perception of objects for Westerners is usually unaffected by field.

Because of the attention difference, evaluators from East Asia and the West may tend to find different usability problems in the usability test. East Asian evaluators and Western evaluators may focus on different parts of the application which may influence the identified usability problems. Evaluators may tend to put more attention on the parts that they are inclined to focus, and ignore other parts that the users mentioned. Further, attention differences may also influence evaluators' understanding of the users' behaviors and speech. The usability test involves both the user's and evaluator's cognitions. The evaluator is not able to follow the user's every thought. Sometimes, although the user wants to emphasize something, the evaluator may not notice it.

Furthermore, evaluators may influence user's behavior because of the different attention focus. A field study (Nørgaard & Hornbæk, 2006) showed that in usability tests, evaluators seemed to seek confirmation of problems that they were already aware of. They often asked users about their expectations and about hypothetical situations, rather than about experienced problems. The field study implies that although the purpose of usability testing is to learn usability problems from users, the evaluator's own preconceptions and expectations may also play an important role in the test. In usability testing, Western and East Asian evaluators may give different responses to the users, which, in turn, may influence the identified usability problem.

Causal Attribution, Prediction and Postdiction

Causal Attribution, Prediction

One of the best established findings in cognitive social psychology is called "correspondence bias" or "fundamental attribution error," which means "the tendency to see behavior as a product of the actor's dispositions and to ignore important situational determinants of the behavior" (Nisbett et al., 2001, p. 298). If it is the case in usability testing, observing a particular user behavior may result in different explanations by different evaluators.

The difference of causal attribution and prediction is originated from holistic and analytic cognitive styles. One of the most important characteristics of the holistic cognitive style is "an orientation to the context or field as a whole" (Nisbett et al., 2001, p. 293). As discussed above, East Asians are more oriented toward contextual factors than are Westerners, so when asking them to explain some events, East Asians will refer to contextual information more than do Westerners. On the other hand, analytic cognitive style emphasizes detaching the object from its context, and focusing on "attributes of the object" (Nisbett et al., 2001, p. 293). When explaining the same events, Westerners, who tend to be analytical thinkers, will focus mainly on the events themselves.

According to Nisbett (2003; 2004), Westerners are inclined to explain events by reference to properties of the object. East Asians are inclined to explain the same events with reference to interactions between the object and the field. For example, Americans tend to explain murders by invoking presumed traits, abilities, or other characteristics of the individual, whereas Chinese tend to explain the same events with reference to contextual factors (Nisbett, 2004). Moreover, when making predictions, East Asians tend to mention "how people in general will be expected to behave," whereas Westerners tend to mention "the behavior of a particular individual" (Nisbett et al., 2001, p. 298). Due to the causal attribution and prediction differences, in a usability testing, Western and East Asian users may give different explanations to their behaviors, which may influence the evaluators' identified usability problems and rated severities. For example, East Asian users may not only talk about their own opinions, but also talk about general opinions or other peoples' feelings. When considering other people's feelings, users' opinions may be not consistent with their actual performance. It may be easier for local evaluators to understand the users' real feelings of the interface, but harder for foreign evaluators to get the users' real feelings. On the other hand, Western users may make predictions of whether other people like the application or not, based mainly on their own feelings. Different causal attribution and prediction may influence the problem finding and severity rating.

Postdiction

Westerners and East Asians have different attitudes towards unanticipated outcomes. The attitudes toward contradiction may influence the usability problem finding and usability problem severity rating in usability testing. Before discussing the influence of postdiction on usability testing, the postdiction differences should be clarified first.

East Asians are good at paying attention to the field, and notice a broad range of factors that may be used to explain an event. According to Nisbett and his (Nisbett et al., 2001), the broad range of factors perhaps explain the event too readily, which means even unanticipated outcomes will not cause surprise to East Asians compared to Westerners. Hindsight bias may be greater for East Asians. Hindsight bias means "the tendency to assume that one knew all along that a given outcome was likely" (Nisbett et al., 2001, p. 299). Choi and Nisbett (2000) investigated surprise towards unanticipated outcomes and contradictions by East Asians and Westerners. The study showed that Korean participants displayed less surprise and greater hindsight bias than American participants did when a target's behavior contradicted their expectations.

In a usability test, evaluators can usually detect usability problems from the user's attitude, such as surprise. For example, the user clicks a button, but an unexpected page is shown.

Western users may be surprised and the surprise could be considered as a signal showing that there is something that needs to be discussed. However, for East Asians, even though they are also surprised to see the unexpected page, they may not feel the surprise as big as do Westerners. East Asian users may easily find an explanation for why it is on a particular page, not the page they expected. However, it does not mean that they like that design. For instance, the East Asian user may say "I thought it was 'replying email.' Now actually it is 'writing email.' But it is ok. The icon of writing email like this is also clear. I am just not so familiar with this software." How will the evaluator react to this sentence? The East Asian evaluators may feel that the design is ok, even though it needs to be changed for novice users. But Western evaluators may have a different understanding of the user's speech, depending on evaluators' expectations. Western evaluators who have the same expectation as the user (it should be "replying email" not "writing email"), may be surprised to see why it is "writing email," and may think it is a big problem. However, Western evaluators, who have no assumption of what it will be beforehand, may not consider it to be a problem because the user does not show much surprise and says it is ok.

Therefore, for tests with East Asian users, it may be more difficult in some situations for evaluators to get the users' real meaning. However, for tests with Western users, it may be easier for evaluators to get the real meaning because of the users' expression of the surprise.

Categorization

Westerns and East Asians tend to categorize objects differently. In this section, first, categorization differences between Westerners and East Asians are introduced, and then the influences of the categorization differences on interface design and UEMs are discussed.

Since East Asians endorse the holistic cognitive style, they are expected to be more capable than are Westerners to detect relationships between the object and the environment, and the relationships among events in the environment. This feature will influence the way of categorizing objects or events. East Asians are inclined to categorize objects according to their relations. On the other hand, Westerners think analytically, so they tend to isolate the object from the field. Categorizing objects for Westerners therefore is usually done according to its properties. Nisbett and his colleagues found that grouping of objects and events tended to be taxonomic for Westerners and relational for East Asians (Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002a; Nisbett et al., 2001). For example, Chinese and American participants were asked to indicate which two of the following three went together: notebook, magazine, pen. Americans tended to put the notebook and the magazine together because both had pages, whereas Chinese tended to put the pen and notebook together because the pen wrote on the notebook (Nisbett, 2004). Moreover, they also found that Americans tended to group objects on the basis of rules, whereas East Asians tended to classify them on the basis of family resemblance.

Further, categorization difference exists in the application's interface. Choong and Rau's studies (Choong, 1996; Choong & Salvendy, 1999; Rau, Choong, & Salvendy, 2004) found that Chinese people tended to classify stimuli according to their thematic relations. Chinese users would have better performance when using a thematic rather than a functional computer system. On the other hand, Americans would have better computer performance when using functional system.

Categorization difference is useful for the card sorting technique. Card sorting is used for exploring how people group items, and thus designers can develop structures that maximize the probability of users finding the items. Because of the categorization difference between East Asians and Westerners, when designing interfaces, the structures, such as the structure of menus, should be organized according to the target users' cognitive styles, characteristics and preference (Nawaz & Clemmensen, 2007).

Categorization difference may also influence usability testing. For example, if an interface structure suits the Westerners' structure, but not the East Asians', it may be easier for East Asian evaluators to detect the problem when users mention it. For the application and target user group with which the evaluator is unfamiliar, the evaluator may need to spend more effort in pursuing the user's real opinions. In order to understand the user's real thoughts, apart from observing and listening, the evaluator also needs to communicate with the user effectively.

Perception of Control

East Asians are situation centered and are sensitive to their environment, whereas Westerners are individual centered and expect their environment to be sensitive to them (Nisbett et al., 2001, p. 295). The different attitudes toward perception of control may influence usability testing.

Psychologist Chiu (1972, cited from Nisbett et al., 2001, p. 295) claims that "Chinese tend to assume a passive attitude while Americans tend to possess an active and conquering attitude in dealing with their environment." Nisbett and Masuda (2003, p. 11163) argue that "frequently lacking explicit rules about the object's behavior, the Chinese had relatively little sense of personal agency or control." From the above statements, we can see that in usability testing, Western users may expect to perceive more control in a given situation, whereas East Asian users may be used to conforming to the situation.

Moreover, the perception of control may also influence the evaluators' way of facilitating tests. In usability testing, Western evaluators may be more willing to interact with users, whereas East Asian evaluators may be more willing to follow the users' speech. The study conducted by Ji et al. (2000) indicates that the perception of control can help Westerners improve performance and increase confidence, whereas it has no influence on East Asians, and sometimes even negative influence because of increased cognitive load from control. From their study, we can infer that in order to understand users' behaviors and find relevant usability problems, Western evaluators may be more active in initiating a communication with the users than are East Asian evaluators.

Logic vs. Dialectics

East Asians tend to hold dialectics of the reasoning, whereas Westerners tend to hold logic of the reasoning. The dimension of logic vs. dialectics may affect the Western and East Asian evaluators' usability problem severity rating behaviors, and it may also influence the understanding between the users and evaluators in usability testing. This section first introduces Nisbett and his colleagues' studies of logic vs. dialectics, and then discusses how the dimension may influence the thinking aloud usability testing.

East Asian society is holistic and attaches importance to social relations, and thus in order to maintain harmony, East Asians are expected "to seek compromise solutions to problems, to prefer arguments based on principles of holism and continuity, and to try to reconcile or transcend seeming contradictions" (Nisbett et al., 2001, p. 296). East Asians have developed dialectic reasoning, which involves "reconciling, transcending, or even accepting apparent contradictions" (Nisbett et al., 2001, p. 294). The goal for dealing with contradictory propositions is to search for the "middle" way between extremes. On the other hand, Westerners are more capable of rejecting "the evidence of the senses when it conflicts with reasons" (Nisbett et al., 2001, p. 294), and they usually "set aside experience in favor of reasoning based on logical rules" (Nisbett et al., 2001, p. 296). It seems necessary for Westerners to know that one side or the other is correct.

Briley, Morris, & Simonson (2000) investigated the consumer choices of East Asians and Westerners, and found that Westerners were more likely to give rule-based justifications and East Asians more likely to give compromise based justifications.

Applying the above discussion to usability testing, we could infer that when rating the severity of the usability problems, East Asians may tend to choose the middle rank, rather than the extreme ranks, whereas Western evaluators may not have such tendency.

Moreover, the understanding between users and evaluators may also be affected by different reasoning. Since East Asian users may be inclined to find the good aspects of a bad application and the bad aspects of a good application, it may be easier for East Asian evaluators to get the users' real points, but harder for Western evaluators to understand. On the other hand, since Westerners prefer to reject one proposition and support the other, in usability testing, Western users may be more critical of the prototype if they do not like it, or if they think they have to be critical since the purpose of the test is to find the aspects to be improved. In contrast, East Asians are used to compromising, when seeing the very critical arguments from users, East Asian evaluators may tend to think of the problem as being more critical than the users' real feelings actually are. However, if users give some small positive arguments of the application, East Asian evaluators may perceive the problems to be less critical than the users' real feelings actually are, since East Asians tend to decrease their confidence in their initial position when presented with a weak argument (Nisbett et al., 2001, p. 302).

Task-Focus vs. Socio-Emotional Orientations

Task-focus and socio-emotional orientations play an important role in the thinking aloud usability test. Before discussing the influence of the orientations on the tests, the concepts are clarified.

One of the important characteristics of holistic society is trying to maintain social harmony. Relations are usually built based on "socio-emotional" orientation concerns. In the analytical Western society, one of the important characteristics is the emphasis on the power of the individual. The relations between each other are usually based on "task" orientation. Sanchez-Burks et al. (2000) propose two kinds of orientations in terms of relational schemas: task-focus orientation and socio-emotional orientation. Relational schemas are "cognitive structures that provide goals and expectations about what can be expected to occur in a given situation, what behaviors are or are not appropriate, and which elements of the situation are important to notice and store in memory" (Sanchez-Burks et al., 2000, p. 7). Relational schemas will influence behavior, preferences and social judgments. Task-focus relational schema means that people's effort is directed towards task-related goals, and "attention is focused on monitoring the extent to which these goals are being accomplished" (Sanchez-Burks et al., 2000, p. 4). Socio-emotional relational schema means that people's effort and attention are directed towards "the interpersonal climate of the situation, and they strive to maintain social harmony" (Sanchez-Burks et al., 2000, p. 4). Typically, Northern European and American cultures are task-focus orientation cultures, and East Asian, Latin, Middle-Eastern and Indian cultures are socio-emotional orientation cultures (Sanchez-Burks et al., 2000). If relational schemas are matched, it will be easier for people to coordinate and communicate. Sanchez-Burks et al. (2000, pp. 8-9) discusses relational schemas, indicating that people prefer to interact and communicate with those who have consistent relational schemas, rather than those with inconsistent schemas.

The different relational schemas of task-focus vs. socio-emotional orientations may influence the thinking aloud usability testing. Usability testing was developed in the West and is now used all over the world. It seems that the usability testing environment is a "culture neutral" environment, where "differences in social traditions are put aside and all effort is focused on the common interest, namely the task" (Sanchez-Burks et al., 2000, p. 37). However, Sanchez-Burks et al. (2000, pp. 37-38) argue that "directing all effort toward task specific goals putting aside social traditions and concerns about interpersonal feelings, reflects a culture-specific way of conducting business rather than a 'culture-neutral' approach." However, usability testing was invented by Westerners and applied in Eastern countries with little change. Yeo's study (1998b) found that some cultural issues needed to be considered when doing usability tests in Malaysia, such as power distance. His research implicates that the way to do usability testing should possibly not be the same in East Asian and Western countries. His study can also be explained by the theory of task-focus vs. socio-emotional orientations. Malaysian culture is a socio-emotional relational orientation culture. If the users consider the evaluator to be of a higher rank, they will be more reluctant to provide negative comments, since they have a socio-emotional orientation and hope to build a good relationship with the higher ranking evaluator. East Asian users who tend to have socio-emotional orientation may be easily influenced by the evaluators' characteristics, status and behaviors, but Western users who tend to hold task-focus orientation may not be influenced by the evaluators.

In this research, both foreign and local evaluators are used to conduct the tests with East Asian and Western users. From the discussion of the task-focus versus socio-emotional orientations, we can say that users from different cultures may be affected by foreign evaluators in quite different degrees. Western users who tend to have a task-focus orientation may not be influenced very much by the evaluators' cultural backgrounds, since they may just focus on their tasks and not care whether the evaluators are native or foreign. On the other hand, East Asian users who tend to hold a socio-emotional orientation may be influenced more easily by the evaluators' cultural backgrounds. In Vatrapu and Pérez-Quiñones (2006)'s study, even in the structured interview, Indian users who tended to have a socio-emotional orientation found more usability problems and made more suggestions to an interviewer who was a member of the same (Indian) culture than to the foreign interviewer.

Discussion: the Potential Influence of Holistic and Analytic Cognitive Styles on the Thinking Aloud Usability Testing

The culture theory developed by Nisbett and his colleagues emphasizes two cognitive styles between East Asians and Westerners, indicating a holistic, dialectical information processing, and an analytical, linear thinking style, respectively (Lehman, Chiu, & Schaller, 2004). The different cognitive styles may have some potential influence on usability testing. On the one hand, evaluators with different cognitive styles may find different usability problems, even though they observe similar user behaviors. On the other hand, evaluators/users with different cognitive styles may have different communication patterns in order to find usability problems. In order to do the usability tests effectively in another culture, before the tests, the usability practitioners should have knowledge of the local users. Actually, the first step in the usability process is to study the intended users and their use of the product (Nielsen, 1993). The culture related information of the target users and the product may be helpful for the usability practitioners to facilitate the tests and find relevant problems. During the usability tests, in order to make sure that there is no misunderstanding of the users' behaviors and speech, evaluators can also write down the unclear things and do more interviews with the users after the tests.

To sum up, we have discussed the understanding between evaluators and users (see logic vs. dialectics in section 2.2.2.2), the way to facilitate the tests (see field independence vs. field dependence, perception of control in section 2.2.2.2), problem finding and severity rating behaviors (see attention to object vs. field, categorization and causal attribution, prediction and postdiction parts in section 2.2.2.2), and the potential influence of relationship building (see task-focus and socio-emotional Orientations in section 2.2.2.2). One further influence needs to be discussed, which is verbalization.

The thinking aloud usability testing is a cognitive process (Hertzum & Jacobsen, 2001). The cognitive difference between East Asians and Westerners implicates that a good usability evaluation method should fit the characteristics of the people in the target country. As mentioned earlier, Westerners' way of thinking is, as characterized by Nisbett, analytical, meaning that they tend to "think in a line" (Clemmensen et al., 2009, p. 213). However, East Asians' way of thinking is seen to be more holistic in that they tend to "think in a circle" (Clemmensen et al., 2009, p. 213).

Kim (2002) states that thinking aloud is best suited to analytic cognitive tasks, while holistic tasks are more difficult to verbalize. A previous study found that verbalization would not interfere with European American participants' cognitive performance, whereas verbalization did interfere with East Asian American participants' performance (Kim, 2002). The reason given by Kim (2002) is that when East Asians, who tend to adopt holistic thinking, wanted to grasp the gestalt of the part, many elements would be held in the thought at the same time, which would make the verbalization more difficult to do. In contrast, Western people, who tended to adopt analytical thinking, broke up the object into component elements which made the verbalization easier to do.

Hence, thinking aloud may not affect the task performance of Westerners who tend to think analytically, but may impair the task performance of East Asians who tend to think holistically. Before doing thinking aloud usability testing in a target country, the usability practitioners need to consider whether the same thinking aloud method can be used in the same way in the local community. If the influence of the local culture on the thinking aloud method is ignored, the results may provide inaccurate information about the localized product. The above discussion implies that the way to do thinking aloud usability test in East Asian and Western countries may not be exactly the same. If thinking aloud does influence the East Asian users' performance, usability practitioners need to consider a better way to do the test and reduce the influence. For example, if task performance is important in the test, the retrospective thinking aloud could be used to collect the verbalization of the user's performance after the performance is over (Guan, Lee, Elisabeth, & Judith, 2006; Haak et al., 2003). On the other hand, if getting the user's opinions (not task performance) is important for the test, then the evaluator's active interaction with the user could be more necessary for tests in East Asia than for tests in the West.

2.2.2.3 Discussion: Cognitive Style and the Usability Testing Model

From the discussion in this chapter, we can see that there is a relation between the cognitive styles and usability testing, see Figure 5.

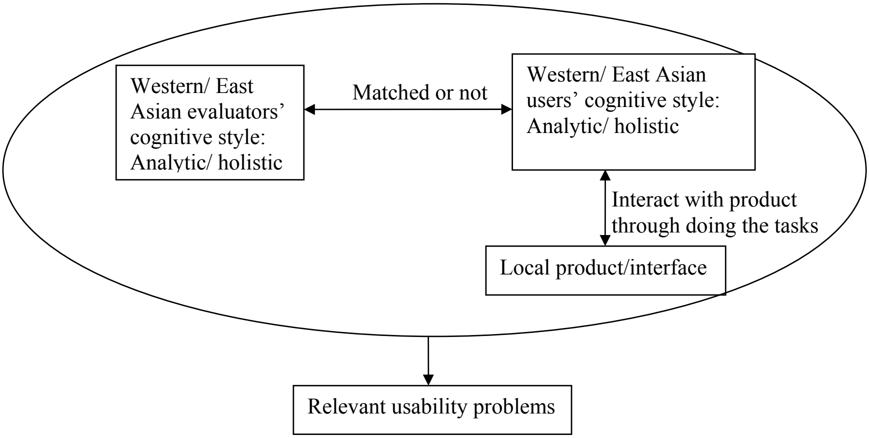

Western/ East Asian evaluators' cognitive style: Analytic/ holistic
Matched or not
Western/ East Asian users' cognitive style: Analytic/ holistic
Interact with product through doing the tasks
Local product/interface
Relevant usability problems

Figure 5: Model of cognitive styles and usability testing

As discussed earlier, usability testing is a cognitive activity (Hertzum & Jacobsen, 2001), involving both the evaluator's and user's cognition and perception. The matching of cognitive styles is important for participants to have a positive attitude towards other people in the same group (Bull & McCalla, 2002; Hayes & Allinson, 1996; Mullany et al., 2007). However, from the discussion in this chapter, we perceive that Westerners and East Asians have different cognitive styles, analytic vs. holistic cognitive styles, which may influence the identified usability problems. If the evaluator's cognitive style and the user's cognitive style are matched, they may feel more comfortable to interact with each other. Further, when cognitive styles are matched, from the cognitive differences discussed above, it may be easier for the evaluator to understand the user's real meaning, and the interaction and communication may be more fluent compared to those whose cognitive styles are not matched. Furthermore, since the user interacts with the local product/interface, local evaluators who tend to have the same cognitive style as the users may have similar feelings and opinions of the product/interface. Hence, the matching of the evaluator's and user's cognitive styles may result in finding more relevant usability problems by evaluators compared to those with different cognitive styles.

The usability problems in Figure 5 are called relevant (or real) problems. A usability problem is relevant if it predicts a problem that users will encounter in real work-contextual usage (Hartson et al., 2001; Wilson, 2007). The concepts of hits, misses, false alarms and correct rejections in signal detection theory are used to describe the usability problem detection by researchers (Gray & Salzman, 1998a; Hartson et al., 2001). According to the signal-detection theory (Maloney, 2003), a relevant usability problem signifies a true attack, not a false alarm. A false alarm is a problem found by evaluator, but not a problem experienced by users (Tullis & Albert, 2008, p. 101). From the above discussion, it is assumed that if the evaluator's cognitive style is matched with that of the user, the evaluator may find more relevant usability problems and fewer false problems, compared to the pairs with unmatched cognitive styles.

Foreign evaluators who have a different cognitive style from that of local users may be strongly influenced by their own cultural perspective, perception and cognition when doing the usability testing. In order to find relevant usability problems, foreign evaluators need to spend extra effort on preparing the tests. For example, before the tests, foreign evaluators need to take more time to learn the localized application in the target culture, compared to testing the application in their own culture. Foreign evaluators need to know the background, user habits and some related cultural features of the application in order to help them understand and communicate with the target users. Furthermore, evaluators should be aware of the target users' way of thinking and talking. For example, East Asian users may not be good as Westerners at thinking aloud, so evaluators should think of a better way to communicate with users.

According to Vygotsky (1962), thinking and speech are related for adults. Wilhem von Humboldt states that "language is the formative organ of thought…Thought and language are …one and inseparable from each other" (Slobin, 1991, p. 70). Language is regarded as an instrument of thought and an instrument of communication (Clark, 1991, p. 325). People from different cultures may have not only different cognitive styles, but also different communication orientations. In the following section, another view of culture---"culture is communication" is introduced (Hall, 1990, p. 94).

2.2.3 Culture is Communication

2.2.3.1 Communication Process in the Thinking Aloud Usability Testing

Before introducing Hall's communication orientations across cultures, the communication model in this research is first clarified. Communication is defined variously by different researchers. For example:

- Communication is defined as the process of using signs and symbols that elicit meanings in another person or persons for whatever intent, or even without conscious intent, on the part of the person producing the symbols or signs (Kim & Gudykunst, 1988, p. 25).
- Human communication is the process through which individuals- in relationships, groups, organizations and societies- respond to and create message to adapt to the environment and one another (Samovar & Porter, 2003, p. 15).

In this study, communication is defined both as the message/ information transmitted from the sender to the receiver, and as the message/ information interpreted by the receiver. The user's main role is the sender, and the evaluator's main role is the receiver. However, since communication is the interaction of two people (Kittler, 2009), the evaluator could also act as the sender and the user could also be the receiver sometimes. For example, when the user does not mention an important issue, the evaluator may probe some questions to elicit the user's opinions on that issue. In this situation, the evaluator is the sender who calls the user's attention to a specific issue, and the user is the receiver to get the evaluator's question. This model could be expressed as Figure 6, which is developed based on Krippendorff's (1986, p. 25) communication process-model.

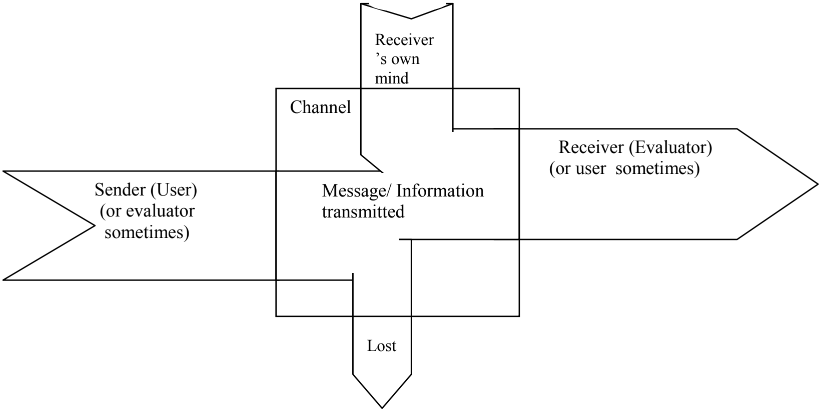


Figure 6: Process-model of communication

Figure 6 shows that the messages/ information sent and received do not have to be exactly the same, though they have to correspond one to another (Kittler, 2009). For the sender, some information is transmitted from the sender to the receiver, and some information may be lost. For the receiver, since he/she uses his/her own mind to interpret the sender’s information, the information received by the receiver may not be exactly the same as that sent by the sender, but processed by the receiver himself/ herself.

The sender is the one who produces the message/ information and the receiver is the one who perceives the message/ information through the communication process channel. In the thinking aloud usability test, normally the user is the sender and the evaluator is the receiver. According to Ericsson and Simon’s model (1993), the only time that the evaluator plays the role of the sender is when the user keeps silent for a while and the evaluator asks him/her to “keep talking.” However, according to Boren and Ramey’s thinking aloud model (Boren & Ramey, 2000), the evaluator should not passively listen, but should actively interact with the user. Tamler (2001) suggests that thinking aloud data which is generated by users themselves is often inadequate. The evaluator needs to probe questions which are important for the interface but not noticed by the user. Moreover, the evaluator also needs to probe questions to ask the users to clarify their meaning (Tamler, 2001). Therefore, the evaluator could be the sender and the user could be the receiver when the message, such as the probing question, is given from the evaluator to the user.

Since there are foreign evaluators in this research, an important concept "intercultural communication" should be introduced.

2.2.3.2 Intercultural Communication in the Thinking Aloud usability testing

Intercultural communication is primarily conceived as "direct, face-to-face communication encounters between or among individuals with differing cultural backgrounds" (Kim & Gudykunst, 1988, p. 12). It occurs "whenever a message produced in one culture must be processed in another culture and can be seen as the intended transport of meaning from one system to another with each system embedded in or socialized by different culture systems" (Kittler, 2009, p. 3). All communication could be regarded as intercultural to some extent, since people hardly share exactly the same culture if the definition of culture is not limited to national or racial groups, but open to all levels of groups (Kim & Gudykunst, 1988; Rogers, Hart, & Milke, 2002). Kim and Gudykunst (1988, p. 13) claim that "the distinction between intercultural and intracultural communication, therefore, is viewed not as a qualitative, categorical distinction, but as a matter of a researcher's particular operationalization of the concepts, or 'drawing of a line' between them." Intracultural communication in this research refers to the communication between people in the same country, and intercultural communication refers to the communication between people from different countries.

When communicating with people from one's own culture, there is reliance on "internalized cultural rules that prescribe appropriate behavior for the communication situation" (Samovar & Porter, 2003, p. 203). People from the same culture can communicate effectively without having to think about those rule (Samovar & Porter, 2003, p. 203). On the other hand, when communicating with people from another culture, one needs to be aware of "how diverse cultural rules influence the communication context" (Samovar & Porter, 2003, p. 203) and consider those rules in the communication, or else there may be unpleasantness or misunderstandings. For example, previous work (Spencer-Oatey, 2000, p. 99) shows that English speakers' best response to a compliment is to accept it, whereas Chinese people's best response is to reject or deny it.

Conversational indirectness is considered as a cause of interpersonal misunderstanding (Sanchez-Burks et al., 2003). Indirectness means "a discrepancy between sentence meaning and speaker meaning," involving what information the speaker sends out and what information the listener gets (Sanchez-Burks et al., 2003, p. 364). Sanchez-Burks, Lee et al. (2003) found that

Chinese and Korean people used more indirectness than European Americans in work settings. This implies that there may be misunderstandings when a Chinese person communicates with an American person at work. For example, even though the Chinese person says that "this study is interesting," the real meaning he intends to convey is that it has many problems. In the usability testing settings, when doing tests with users in different cultures, evaluators should be aware of people's preferred ways of expression.

Communication is a way of transmitting information, and all meaningful information is surrounded with context (Kittler, 2009). In intercultural communication studies, context is an important issue substantially discussed by many researchers (Hall, 1989a, 1989b; Hall & Hall, 1990; Kittler, 2009; Luzio et al., 2001; Rogers et al., 2002; Samovar & Porter, 2003). This study utilizes Hall's high- and low- context communication orientations, which are introduced below.

2.2.3.3 Hall's Concepts of High-Context and Low-Context Communications

Hall (1989b, p. 60) claims that "the matter of context will be an issue in any communication between human beings," and "no communication is totally independent of context, and all meaning has an important contextual component." According to Hall 1976 (cited from Hall & Hall, 1990, p. 6), a high context communication or message is "one in which most of the information is already in the person, while very little is in the coded, explicit, transmitted part of the message" (Hall & Hall, 1990, p. 6). A low context communication is "just the opposite; i.e., the mass of the information is vested in the explicit code" (Hall & Hall, 1990, p. 6).

An example of high context communication given by Hall (1989b, p. 60) is that a man and a woman have lived together for more than fifteen years and they can communicate more economically without spelling out everything. An example of low context communication (Hall & Hall, 1990, p. 6) is two lawyers in a courtroom during a trial and they have to spell out everything. Figure 7 shows the relation between the high and low context, transmitted information and meaning.

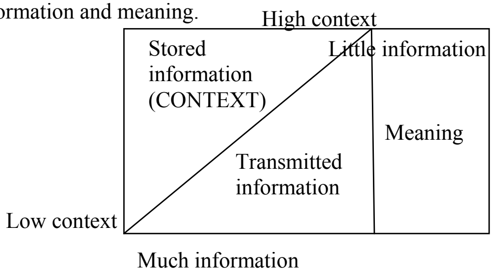


Figure 7: Context and transmitted information in order to understand the meaning (Hall 1989, p. 61).

Hall (1989b, p. 60) states that "information, context and meaning are bound together in a balanced, functional relationship." Figure 7 is developed by Hall, which shows that in order to make the meaning remain constant, if context is lost, information must be added (Hall, 1989b, p. 61). When a person from a high-context culture communicates with the one who is from a low-context culture, there may be some problems. According to Hall and Hall (1990, p. 9), "high-context people are apt to become impatient and irritated when low-context people insist on giving them information they do not need. Conversely, low-context people are at a loss when high-context people do not provide enough information." Thus, in order to make the communication easier, the appropriate amount of information and the level of context should be kept in mind for the intercultural communication.

High-Context and Low-Context Communication in the Thinking Aloud Usability Testing

Americans, Germans, Swiss, Scandinavians and other northern Europeans are regarded as low-context people, whereas East Asians, Arabs and Mediterranean people are regarded as high-context people (Hall, 1989a, 1989b; Hall & Hall, 1990). This section takes Danish and Chinese thinking aloud usability testing settings as examples to discuss the thinking aloud usability tests in the low-context and high-context cultures. Danish evaluators and users are from a low-context culture, whereas Chinese evaluators and users are from a high-context culture. In usability testing, when the communication involves Danish people, they expect to have a lot of information and many details from a partner, whereas when the communication involves Chinese people, they expect the partner to know something and therefore detailed information is not necessary to give.

As discussed above, in usability testing, both the user and evaluator could be the sender and receiver, even though the user is the main sender and the evaluator is the main receiver. Based on the six communication archetypes proposed by Kittler (2009), this research has developed the intercultural communication model for the Danish users and Chinese evaluators (Figure 8), as well as the communication model for the Chinese users and Danish evaluators (Figure 9).

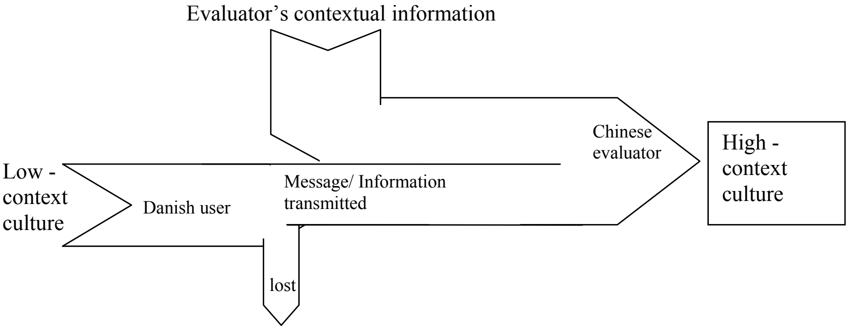


Figure 8: Information flow from the Danish user to the Chinese evaluator

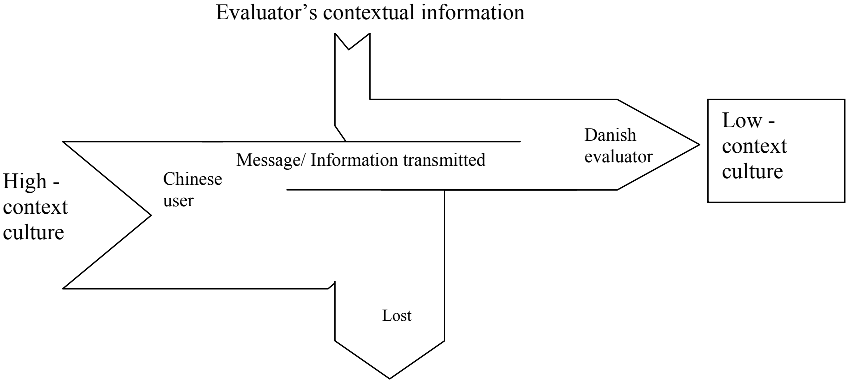


Figure 9: Information flow from the Chinese user to the Danish evaluator

From Figure 8, we see that since Danish users are low context orientation, they tend to give information without much context information lost. However, Chinese evaluators are high-context orientation and tend to get more information beyond the "pure" information sent by users. The message received may be distorted by the additional assumptions and interpretations of evaluators' own contextual knowledge. Some information may thus not be the Danish users' intended meaning, but the evaluators' own thoughts. The evaluators' and users' meaning may correspond only to a medium degree.

Figure 9 shows the information transmission from the Chinese users to the Danish evaluators. When sending a message, Chinese users may tend to emphasize context-bound information, whereas Danish evaluators tend to focus on the context-free information contained in the message. Some of the information sent by Chinese users may be lost to the Danish evaluators, since the context information may not be recognized by the evaluators. The meanings also correspond only to a medium degree.

The above discussion is about the relevance of intercultural communication for this research. Of course, the intercultural communication also includes the information flow from the evaluator to the user. Figure 8 can be seen as the information from Danish evaluators to Chinese users, and Figure 9 can be seen as the information from the Chinese evaluators to Danish users. The explanation is similar to that of "user to evaluator" discussed above. Apart from intercultural communication, the research also involves intracultural communication, namely, the tests done by the local pairs. Since the local pairs are from the same culture with the same context orientation, Danish evaluators may know that what the user has said is what the user meant, and Chinese evaluators may know that what the user said makes best sense in the particular context. Chinese users tend to talk around a point and expect the evaluator to understand what they mean, whereas Danish users take the main responsibility to convey the meaning.

2.2.3.4 Discussion

In these sections, culture is regarded as communication, which means people with different cultural backgrounds may have different communication orientations. As cognitive styles, the communication orientation can also be stable over several years (Levy, Wubbels, Brekelmans, & Morganfield, 1997). Since communication is very important in the thinking aloud usability testing for evaluators to understand users and come up with usability problems, the evaluators' and users' communication orientations cannot be ignored.

In the test, evaluators and users with different communication orientations may show different communication patterns. The identified usability problems may also be impacted by their communication orientations. The thinking aloud method is an open-ended method with little structure (Norman & Panizzi, 2006), which means that there is little "multiple-choice" questions or checklists because "adding structured questions might lead the user or suggest answers that they would not otherwise have made" (Norman & Panizzi, 2006, p. 251). But usually open-ended questions result in qualitative data which need to be interpreted and categorized by evaluators (Norman & Panizzi, 2006, p. 251). Since the test is not highly structured, evaluators from different cultures may interact with users differently because of the different communication orientations and cognitive styles, which may result in different usability problems and communication patterns.

Communication Model in the Thinking Aloud Usability Testing

In order to further understand the communication patterns in the thinking aloud usability testing, a communication model is developed. Figure 10 shows the communication model between the evaluator and the user, and also between the user and the application in usability testing.

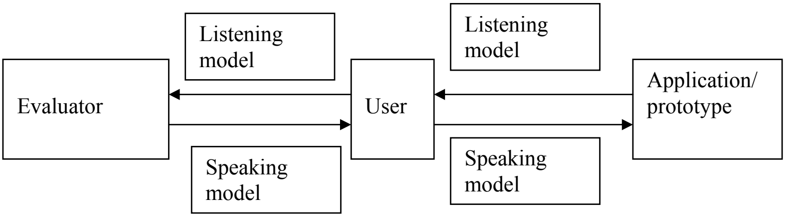


Figure 10: Communication model in the thinking aloud usability testing

The communication model in Figure 10 is developed based on the listening and speaking model for the user and computer interaction, which is produced by Ito and Nakakoji (1996). Many previous researchers who considered the cultural issue in HCI area investigated the interaction between the user and the interface, such as how to design an interface suitable for a specific cultural group (Bourges-Waldegg & Scrivener, 1998; Bourges-Waldegg & Scrivener, 2000; Callahan, 2004; Choong, 1996; Marcus, 2005; Shen et al., 2006). Ito and Nakakoji (1996) developed a model of user-computer interaction and considered the way that culture influenced each step. The model includes two modes: listening mode and speaking mode. Interaction happens in these two modes: "*listening* mode, in which people are presented with a computer's reaction, and *speaking* mode, in which people give instructions to a computer system" (Ito & Nakakoji, 1996, p. 108). Because of the purpose of this research, this study discusses only the communication between the evaluator and user, not the user and the application.

The listening mode and speaking mode can also be used to discuss evaluator’s and user’s interaction. Turning now to the evaluator’s perspective, the evaluator is the one who facilitates the test and comes up with usability problems, and the purpose of the research is to develop some guidelines for evaluators to do tests in different cultures. The listening mode refers to that which the evaluator is presented with the interaction between the user and interface, including both observation and listening; in the speaking mode, the evaluator interacts with the user, such as giving instructions, probing questions, etc.

In the listening mode, the evaluator’s cognitive activity can go through the three phases which are proposed by Ito and Nakakoji (1996) in order to comprehend the presented information: perception, semantic association and logical reasoning. The perception phase here refers mainly to the sensation which is “least affected by culture” (Callahan, 2004, p. 271) because the physiology of sound or gesture depends mainly on sensory monitors. In the semantic association phase, the user associates meaning with text and graphics of the interface and speaks out. The evaluator based on his/her own knowledge/cognition system, associates meaning with the user’s speech. Culture plays a larger role in this phase because it needs the evaluator’s understanding of the user’s speech, which involves the user’s verbalization and the evaluator’s information processing. The reasoning phase involves the evaluator’s cognitive reasoning process which is most affected by culture. The evaluator’s cognitive reasoning depends on his/her background culture, knowledge and way of thinking. From cognitive reasoning, the evaluator will have a deep understanding of the whole testing and come up with usability problems. The listening mode cannot be separated from the speaking mode, since the evaluator needs to speak with the users to understand their real meaning and confirm his/her understanding of the users’ interaction with the application/ prototype.

In the speaking mode of the user and the application/ prototype interaction process, users convey their intentions to the computer (Ito & Nakakoji, 1996, p. 112). In the speaking mode of the evaluator and user interaction process, the evaluator expresses his/ her understanding of the users’ behavior and makes clear the ambiguous meaning produced by the users. Usually, in the thinking aloud usability testing, the user is the main talker and the evaluator only talks when necessary, not all the time. The necessary occasions usually are those when figuring out the misunderstandings or getting more deep information. As the communication theory of interaction suggests: “The pattern of much human interaction reflects the tendency of man to behave in ways that minimize the internal inconsistencies among his beliefs, feelings, actions, and interpersonal relations" (Sereno & Mortensen, 1970, p. 178). When the user's behavior is different from the evaluator's expectation or when the user's thinking aloud is not clear, usually the evaluator will interact with the user to figure it out. In the speaking mode of the evaluators' and users' interaction, there could also be three phases: questions emerging, applicability check and enactment with expectations (Callahan, 2004; Ito & Nakakoji, 1996).

- In the question emerging phase, the evaluators identify what further information they need to get and what they need the users to clarify. Since users do not spontaneously address everything of importance, skilled evaluators need to probe questions to call the user's attention to facts not noticed or commented on, etc (Tamler, 2003). Moreover, thinking aloud data is always unclear or ambiguous, and thus it is difficult for evaluators to understand sometimes, especially when evaluators and users have different cognitive styles and communication orientations. Thus, evaluators need to probe questions to ask the users to clarify.
- The second phase is the applicability check. In this phase, the evaluators need to consider: 1) when it is suitable for them to talk, such as when the users stop talking or when the users finish the sub-task and have a break; 2) what actions they are going to take, such as probing questions or sharing the interpretations and getting the user's feedback. Culture may affect this phase. For example, evaluators with different cultural backgrounds may have different preferences in probing questions. Evaluators with low-context communication may probe more questions to users with high-context communication than they would to users from their own culture, whereas evaluators with high-context communication may not have such tendency.
- Enactment with expectations: the evaluator will carry out the action. This phase may also be influenced by the participants' cultural backgrounds. For example, evaluators with task-focus orientation and low-contextual communication may probe questions directly, whereas evaluators with socio-emotional orientation and high-contextual communication may probe questions indirectly.

2.2.4 Summary

In this study, the potential influences of different cognitive styles and communication orientations on the thinking aloud usability testing are discussed. Culture is defined as both cognitive style and communication in the PhD thesis. Cognitive styles and communication orientations are not separated from each other, but connected. The high-context and low-context communication orientations can be related to the holistic thinking and analytic thinking styles. High-context communication orientation is close to the holistic thinking style in considering the context as a whole, and it is hard to separate the context/ field to the object, whereas low-context communication orientation is close to analytic thinking style in de-contextualizing the object from the context/ field. Also, high-context communication is more likely to be used in situations "in which social relations are important," whereas low-context communication is often used in situations "where social relationships are not so important but the task at hand is" (Kimmel, 2000, p. 633). This is close to the task-focus vs. socio-emotional orientations in Nisbett's theory. Moreover, from the linguistic point of view, language plays a cultural role in the development of both cognitive styles and communication orientations (Clark, 1991; Clemmensen, Hertzum et al., 2008; Kimmel, 2000; Vygotsky, 1962). Hence, the culture theory of holistic- and analytic-cognitive styles and the communication model of the high- and low- context communication orientations are two sides of the same coin. Different cognitive styles and communication orientations may influence the thinking aloud usability testing from the theoretical point of view. In the following section, the research question and sub-research questions are discussed, and the hypotheses and themes based on the theories are put forward.

2.3 Research Question, Hypotheses and Themes

In this study, the research question is investigated through two sub-research questions (section 1.2). The two sub-research questions are investigated by examining some hypotheses and themes. In this section, the research question and sub-research questions are discussed first. Then the hypotheses and themes which are investigated in this study are put forward.

2.3.1 Research Question

As discussed in section 2.2, culture may impact many aspects of the thinking aloud usability testing. The research question is "to what extent does the evaluators' and users' cultural background influence the thinking aloud usability testing.” In this PhD research, the main focus is on investigating culture’s impact on usability problems, and evaluators’ and users’ communications. They are presented as two-sub research questions:

R1. In the thinking aloud usability testing, to what extent does the evaluators’ and users’ cultural background influence the evaluators’ usability problem finding and usability problem severity rating?

R2. To what extent does the evaluators’ and users’ cultural background influence the evaluators’ and users’ communications in the thinking aloud usability testing?

As discussed in section 1.1 and 1.2, usability problems could be regarded as the result of the thinking aloud usability testing. Comparing the problems identified by local and foreign evaluators could be regarded as a way to examine the impact of different cognitive styles and communication orientations on usability testing. Apart from the identified usability problems, communications which could be regarded as the process of the usability testing may also be influenced by the evaluators’ and users’ cognitive styles and communication orientations. Evaluators and users with different cultural backgrounds may have different patterns of communications in order to understand each other and find the relevant usability problems. In the next sections, the usability problems and communication patterns are discussed separately.

2.3.1.1 First sub-research question: Usability Problems

Identifying usability problems is the main purpose for doing usability evaluation (Hartson et al., 2001). Usability problem analysis, such as problem counting, problem matching or problem severity rating, is often used as the main approach to assess different UJEMs or different conditions of using a particular UEM (Als, Jensen, & Skov, 2005; Blackmon, Kitajima, & Polson, 2005; Hartson et al., 2001; Hertzum & Jacobsen, 2001; Hornbæk, 2009; Hvannberg et al., 2007; Kjeldskov & Stage, 2004; Skov & Stage, 2005; Virzi et al., 1996). For example, Virzi et al. (1996) compared the identified usability problems to see whether low-fidelity prototypes could be as effective as high-fidelity prototypes in the design process. In this research, in order to examine to what extent the evaluators’ and users’ cultural backgrounds influence the thinking aloud usability testing, comparing local and foreign evaluators’ identified usability problems is a common and acceptable approach.

Usability problems can be divided into two criteria: problem severity and problems that detected by single or more groups of evaluators (Hertzum, 2006; Law & Hvanneberg, 2004; Molich, Ede, Kaasgaard, & Karyukin, 2004). Moreover, in the cross-cultural studies, culturally determined usability problems are proposed by researchers (Bourges-Waldegg & Scrivener, 1998). Furthermore, in order to better understand the usability problem finding and problem severity rating behaviors in different cultural settings, an important concept "evaluator effect" is worthwhile introducing. In the following sub-sections, important issues are clarified.

Usability Problem Severity

The main purpose for rating the problems' severity is to prioritize them and solve the most severe problems first, since it is impossible to solve all the problems because of the limitation of time and money (Hassenzahl, 2000). Hertzum (2006) found that half of the severity sum (problem impact) was concentrated in approximately 20% of the problems, which meant that if 20% of the most severe usability problems were solved, the application's assessment would be improved 50%.

A common way to estimate the usability problem severity is the experts' judgements which can be done by asking the usability specialists to rate the severity of each problem. Using a single scale is a common approach to rating the usability severity (Nielsen, 1993, p. 103). An example of a single rating scale for the severity of usability problems is (Nielsen, 1993, p. 103):

0= this is not a usability problem at all

1= cosmetic problem only- need not be fixed unless extra time is available on project

2= minor usability problem- fixing this should be given low priority

3= major usability problem- important to fix, so should be given high priority

4= usability catastrophe- imperative to fix this before the product can be released

Evaluators make a single rating on a severity scale as shown above. Hertzum and Jacobsen (2001) found that evaluators had different strategies for assessing the severity of problems. This study investigates whether evaluators with different cultural backgrounds have different preferences in rating the usability problem severity.

Further, experts' judgment in rating the problem severity can be regarded as a subjective measure (Hassenzahl, 2000). The usability problem severity could also be estimated by empirical data, such as the impact of a problem (Hassenzahl, 2000). The data-driven approach for estimating severity is an objective measure of severity. Hassenzahl (2000) used error handling time as the objective measure of problem severity. However, he found a fundamental lack of correspondence between the objective measure of severity- error handling time, and the subjective measure of severity- evaluators' judgment. In his study, he neither considered the cultural issues, nor the way of finding usability problems. Therefore, in this PhD research, apart from examining whether evaluators with different cultural backgrounds have different tendencies in giving specific ranks to the problems, the relation between the time on problems and the problem severities will also be investigated by considering the cultural issues and the different ways of finding usability problems.

Shared vs. Unique Usability Problems

Usability problems can be divided according to the problems detected by single or more groups of evaluators, which are referred to as unique vs. shared usability problems, respectively. Unique usability problems are those identified by only one group of usability tests, and shared usability problems are those detected by more than one group of usability tests (Law & Hvanneberg, 2004). The division of shared and unique problems is often used in analyzing problem finding between different groups of evaluators.

Since shared vs. unique usability problems are divided by the differences in evaluators' problem detection (Hertzum & Jacobsen, 2001), the division can also be used to examine the agreement of the usability problems detected by evaluators in one group. If a problem is found by most evaluators, it may be a common problem with higher priority, which needs to be fixed by the designers. In a group of homogeneous evaluators, if a problem is found by only one evaluator in a single test, it might be a critical problem (Molich et al., 2004), but there is a big chance that it is not important. On the other hand, if there are two groups of two different types of evaluators (such as local and foreign evaluators), and a unique problem is found by one group of evaluators, not the other, then the reason behind it may be worthwhile discussing. In this study, shared and unique usability problems will be analyzed between foreign and local evaluators.

Usability Problem related to Culture

Bourges-Waldegg & Scrivener (1998) have proposed culturally determined usability problems, which are those caused by culture or rooted in culturally specific contexts. For culturally determined usability problems, the representations may not be understood by users from other cultures. For example, color, humor, data expression, totems, icons, symbols, pictures, time formats, jargon and abbreviations, are all representations which may vary between cultures due to cultural factors such as language, taste and religion. All these can cause usability problems in the interface shared by people from different cultures. In usability tests, when testing localized application, the foreign evaluators need to know the application well, or else, there will be a bigger chance for them to come up with unreal usability problems and miss real usability problems because of the different knowledge of the application from the users.

Furthermore, people from different places may tend to have different concepts of usability. The different comprehension of usability may make them pay attention to different parts of the application, which may influence the problem finding and problem severity rating. A study about usability construct shows that Chinese participants used constructs related to security, task types, training, and system issues, whereas Danish participants made more use of constructs traditionally associated with usability (e.g., easy-to-use, intuitive, and liked) (Hertzum et al., 2007). If Chinese people are inclined to emphasize the usability of security, task types, training, and system issues of a software/application, they might find more usability problems related to that. On the other hand, Danish people might find more problems with regard to "easy to use" and satisfaction. Thus, this study investigates whether evaluators with different cultural backgrounds find different kinds of usability problems.

Evaluator Effect

The total number of usability problems found will depend upon the knowledge, the experience and the number of the evaluators, which is called the evaluator effect (Clemmensen & Goyal, 2005). Hertzum and Jacobsen (2001, p. 422) define evaluator effect as "the differences in evaluators' problem detection and severity ratings."

Hertzum and Jacobsen (2001) examined three of the most widely used UEMs, cognitive walkthrough, heuristic evaluation, and thinking aloud, and found that all of them suffered from a substantial evaluator effect. No two evaluators evaluating the same interface and using the same UEM found the same set of problems. The evaluator effect exists "for both novice and experienced evaluators, for both cosmetic and severe problems, for both problem detection and severity assessment, and for evaluations of both simple and complex systems" (Hertzum &

Jacobsen, 2001, p. 421). Hertzum and Jacobsen (2001) summarized three potential reasons why different evaluators find different usability problems: a) vague goal analyses; b) vague evaluation procedures and c) vague problem criteria.

In the cross-cultural usability studies, apart from the three potential reasons that Hertzum and Jacobsen (2001) gave, one more reason may be the evaluators' cultural backgrounds. Evaluators with different cognitive styles and communication orientations may pay attention to different issues in the test or may have different understandings of the users' task performing behaviors or speech. Thus, although they may follow the same procedures, use the same task scenarios, have the same understanding of usability, and conduct the tests with similar target users, they may tend to find different problems or rate the severity of the problems differently. This study examines whether the local and foreign evaluators find similar usability problems for culturally localized applications. The results of this study should provide a better understanding of the evaluator effect in usability testing.

Summary

Usability problem analysis is the main method used to assess different conditions of using a particular UEM; thus, it is appropriate to compare local and foreign evaluators' problem finding and problem severity rating behaviors in the thinking aloud usability testing. Section 2.3.1.1 has discussed some important issues related to usability problems in order to have a better understanding of the first sub-research question. The important issues are: usability problem severity, shared vs. unique problems, usability problems related to culture, and evaluator effect. From this research, we hope to understand the usability problem finding and usability problem severity rating behaviors in different cultural settings. Moreover, this research can provide inspiration for interesting topics in the usability research, such as the relation between the objective measure and the subjective measure of the usability problem severity and the evaluator effect.

2.3.1.2 Second Sub-Research Question: Usability Problem Communications

In this research, apart from investigating the cultural impact on usability problems, the cultural impact on the process of testing—communication is also explored. In usability testing, the users speak out their thoughts and the evaluators communicate with users on necessary occasions in order to find the problems. In the usability practice, the usability practitioners do not passively listen but actively interact with users (Boren & Ramey, 2000; Nørgaard & Hornbæk, 2006; Shi, 2008a; Tamler, 2003), and thus communication tends to be very important in the tests. This study involves both intra- and inter-cultural communications. Evaluators and users with different cultural backgrounds may communicate differently in the tests. Moreover, the communication between the distant pairs (users with foreign evaluators) may be different from that of the local pairs, especially when some specific cultural issues are involved in the testing application (Bourges-Waldegg & Scrivener, 1998; Shen et al., 2006; Vatrapu & Pérez-Quiñones, 2006). In section 1.6, some previous usability studies on communication differences in different cultural settings have been introduced. The following section focuses on the communication analysis in this research.

Analysis of the Communication

The communication analyzed in this research includes the communication in tests and the communication of usability problems. The purpose of usability testing is to find usability problems (Nielsen, 1993), making the communication of the problems worthy of investigation. In this study, in order to analyze the communication, conversation analysis (Luzio et al., 2001; Sacks, 1994; Sacks, Schegloff, & Jefferson, 1974) is used to identify the communicative patterns or genres of evaluators and users. Conversation analysis (CA) is started by "analyzing the mechanisms of communicative interaction, especially with respect to the organization of turn taking in conversations" (Luzio et al., 2001, pp. 9-10). CA refers to the exchange of utterances, such as "conversation" or "talk in interaction" (Luzio et al., 2001, p. 10).

In this research, the evaluators' and users' verbal behaviors are coded as point events to analyze the communicative patterns/ genres. Point event behavior (Observer, 2005, p. 92) is behavior which occurs at a point of time, with no duration, but just a timeless spot. Communicative patterns or genres are "relative stable types of utterances" (Boren & Ramey, 2000, p. 267), such as sayings, narratives, greetings. In the thinking aloud usability testing, according to Boren and Ramey's model (2000), the evaluators' verbal behaviors include affirmative responses, non-classic reminders, help behaviors and probing behaviors.

Affirmative response can be used as a sign to show that evaluators are listening and can be seen as a token to encourage the user to continue, such as "ok" "yeah" "huh." The reminder plays the role of reminding the user to keep talking. Ericsson and Simon (1993) proposed “keep talking” as a classic reminder. Boren and Ramey (2000) found that evaluators in practice seldom used the classic reminders in usability tests. For example, they mentioned “and now?” as the reminder to get the user’s to talk immediately without backtracking. This study examines the kinds of reminder that usability professionals prefer to use in the thinking aloud test. Thus, both the non-classic reminder and the classic reminder “keep talking” are coded. Help behavior is also necessary for diagnostic tests with unfinished application (Boren & Ramey, 2000). When the user is stuck with the usability problems, sometimes evaluators need to provide help for the tasks that rely on the user continuing. The main purpose for the evaluator is to find usability problems. If users cannot do the next tasks because those tasks rely on the previous unsolved task, the evaluator needs to help the users in order to make them continue the test. The unsolved task could be recorded as a critical or important problem. The help behavior does not cover the problems, and is coded as one of the evaluator’s communication patterns.

According to previous studies (Boren & Ramey, 2000; Dumas & Loring, 2008; Tamler, 2003), probing behavior is very important in the test because evaluators need to ask the users to clarify their comments or to get information the users have ignored. There are two main kinds of probing behaviors: one is “digging deeper probing” which includes asking for clarification, the other is “directed probing” which is used for getting the users’ opinions/ feelings for specific parts of the application that may be ignored by users. The division of the two probing behaviors is to see whether the probing question is followed with the user’s speech or started by the evaluator. Digging deeper probing is used to ask the user to clarify, which follows the user’s speech, and is the same topic as the user’s current speech content. On the other hand, directed probing is used to start another topic by evaluators, moving further away from Ericsson and Simon’s thinking aloud model (Boren & Ramey, 2000, p. 275). As Tamler (2003) claims, since users do not spontaneously address every important thing, evaluators need to use questions to call users’ attention to the important issues that they are not noticing. Based on the discussion of the high-context and low-context culture, Danish evaluators may give more digging deeper probes than Chinese evaluators, since Danish evaluators are low-contextual communication orientation and need to know more information in order to understand the users’ meaning. For directed probing, evaluators with local users may give more directed probes than those with foreign users, since the evaluators who share the same cognition and perception of the application as that of the users, may have more confidence to probe directed questions to draw the users' attention to the specific part of the application.

So in this research, the evaluators' point events are: affirmative responses, directed probes, digging deeper probes, non-classic reminders, classic reminders, help behaviors. Based on the study done by Vatrapu and Pérez-Quiñones (2006), the user's point event could be coded as: negative comments, positive comments, suggestions and culture related comments.

Apart from the point events of the evaluators' and users' behaviors, this study also analyzes the state event behaviors which show the users' states, "thinking aloud," "talking rather than thinking aloud" and "silence." State event behavior (Observer, 2005, p. 92) is the behavior with duration of time, and with both a start and an end. The point events and state events are not in conflict. There could be point event behaviors in the state events. Point event behavior is used to analyze the action sequences of the evaluators and users to see the communicative patterns/ genres. State event analysis could help in examining the communication in the thinking aloud usability tests. According to the classic thinking aloud model (Ericsson & Simon, 1993), there should be little interaction and communication, but based on Boren and Ramey's model (2000), communication is necessary in a usability test. Thinking aloud is doing the task while verbalizing the thought. Some of the users' verbalization could not be considered as thinking aloud data, but talk, such as comments. In the state of "talking rather than thinking aloud," users tend to give comments without doing tasks or chat with the evaluators. The classic thinking aloud state event could be regarded as the communication pattern of "active user and passive evaluator." Silence could be used to examine the influence of high-context and low-context culture on the thinking aloud test, since silence depends on more context than talk (Andersen, 2001, p. 61). Users from a high context culture may not be willing to talk as much as those who are from a low context culture. The state event behaviors are useful in analyzing the communication patterns. How to code state events will be discussed in detail in the section 3.6.3.

In a word, the communication to be analyzed is based on the action sequence of the evaluators and users. The evaluators' and users' verbal behaviors are coded as point events. The users' state behaviors are also analyzed in order to have a better understanding of the communication patterns and the thinking aloud models.

Summary

In order to propose the themes related to the communications in the next section, the potential communication patterns have been discussed. The codes used to analyze the communications have been put forward based on the thinking aloud models and previous research. In this research, communication patterns in different cultural settings are investigated.

2.3.2 Hypotheses and Themes

Hypotheses and themes are proposed by considering the sub-research questions.

R1. In the thinking aloud usability testing, to what extent does the evaluators' and users' cultural background influence the evaluators' usability problem finding and usability problem severity rating?

R2. To what extent does the evaluators' and users' cultural background influence the evaluators' and users' communications in the thinking aloud usability testing?

The hypotheses and themes are developed based on the theories discussed in this chapter. For the first sub-research question, some hypotheses are put forward. The second sub-research question which is about the communication patterns is a more explorative study; it thus becomes more difficult to have very clear hypotheses beforehand. Instead, some themes are proposed. Themes are the topics which are worthy of investigation in a research study (Frandsen-Thorlacius et al., 2009). The hypotheses and themes are discussed in the following sub-sections.

2.3.2.1 Hypotheses for sub-research question 1

In order to investigate the first sub-question, the hypotheses are developed based on each country. Both Ericsson and Simon's thinking aloud model (1993), and Boren and Ramey's thinking aloud model (2000) show that evaluators should learn the problems from the users. For example, according to Boren and Ramey, test users the experts and evaluators are the learners (section 2.1.3.2). Thus, it could be assumed that usability problems are mainly gained from the users' verbal and non-verbal behaviors. Moreover, the testing subject (section 2.1.3.2) in thinking aloud usability testing is the application, and in this study the localized applications in Denmark and China are different. We could see that both users and applications are different in the two countries, and thus it is not necessary to compare the usability problems across Denmark and

China. The identified problems are only compared within each country. When comparing usability problems brought forward by local and foreign evaluators in one country, the main difference is the evaluators' cultural background. If the identified problems or severity ranks are different, it can be inferred that the difference is from the evaluators' cultural background. For the first sub-research question, the focus is on the evaluators' usability problem finding and severity rating behaviors, not the users'; accordingly, it is appropriate to investigate the sub-research question based on the tests within each country. The hypotheses are:

Hypothesis 1: The usability problems found by local evaluators are different from the usability problems found by foreign evaluators.

H1 a (in Denmark): When with Danish users, the usability problems found by Danish evaluators are different from the usability problems found by Chinese evaluators.

H1 b (in China): When with Chinese users, the usability problems found by Chinese evaluators are different from the usability problems found by Danish evaluators.

Because of the different cognitive styles and communication orientations, the local and foreign evaluators may find different usability problems (section 2.2.2 and section 2.2.3). Hypothesis 1 is divided into two sub-hypotheses since the situation may not be exactly the same as those in Denmark and China. From the task-focus vs. socio-emotional orientations (section 2.2.2.2), we can assume that when doing tests with Chinese users, the difference between the local and foreign evaluators is greater than that with Danish users. Further, when people from high- and low- context communication cultures communicate with each other, the sender's intended meaning may be transmitted only to a medium degree to the receiver (section 2.2.3.3). However, in usability testing, it may be much more difficult for foreign evaluators to get the meaning from users with high-context communication orientation, compared to that of users with low-contextual orientation. Danish users who are low contextual communication try to verbalize everything in their mind, which gives foreign evaluators sufficient information to detect the usability problems and decide the severity of the problems. In contrast, Chinese users who are high contextual communication may not talk about as much, since some information is supposedly known by evaluators. It may be more difficult for foreign evaluators to get the Chinese users' real meaning. Hence, hypothesis 1 is divided into two sub-hypotheses which need to be tested separately. For the same reason, the hypotheses 2, 3 and 4 are also divided into two sub-hypotheses.

Hypothesis 2: Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

H2 a (in Denmark): When with Danish users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

H2 b (in China): When with Chinese users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

Hypothesis 2 is not presented as local and foreign evaluators, but as Chinese and Danish evaluators. H2a and H2b are also based on each country. H2 could also be expressed as local and foreign evaluators. The reason for not doing so is that there is a clear assumption that Chinese evaluators are different from Danish evaluators in rating usability problem severity. According to Nisbett and his colleagues, Chinese people have the tendency of choosing the "middle" solution (section 2.2.2), and thus Chinese evaluators may tend to give the middle rank "important," rather than critical or minor, whereas Danish evaluators may not have such tendency.

Hypothesis 3: The usability problems found by the local evaluators are more consistent with the usability problems found by the target users, compared to the usability problems found by foreign evaluators and target users.

H3 a (in Denmark): The usability problems found by Danish evaluators are more consistent with the usability problems found by Danish users, compared to the usability problems found by Chinese evaluators and Danish users.

H3 b (in China): The usability problems found by Chinese evaluators are more consistent with the usability problems found by Chinese users, compared to the usability problems found by Danish evaluators and Chinese users.

In the thinking aloud usability testing, evaluators should learn the application from the users. Hartson et al. (2001, p. 387) stated that "usability is ultimately determined by the end user, not an expert evaluator, so realness of problems needs to be established by the user." Hence, for the usability problems found by the evaluators, it is meaningful to examine whether users agree or not. In a cross-cultural study, where the local evaluators come from the same culture as do the users, from the high- and low- context communication theories discussed in chapter two (section 2.2.3.3), information transmitted from users to local evaluators may be more than that given to foreign evaluators. Further, since the application in this research is culturally localized (section

3.3.1.1), local evaluators and users may have the same perception of the application (section 2.2.2.2), which may give more opportunity for the local pairs to find similar usability problems.

Hypothesis 4: The usability problem severity ranks given by the local evaluators and target users are more consistent than the problem severity ranks given by foreign evaluators and target users.
H4 a (in Denmark): The usability problem severity ranks given by Danish evaluators and Danish users are more consistent than the problem severity ranks given by Chinese evaluators and Danish users;
H4 b (in China): The usability problem severity ranks given by Chinese evaluators and Chinese users are more consistent than the problem severity ranks given by the Danish evaluators and Chinese users.

Apart from the usability problem findings, the local pairs may also give similar severity ranks for usability problems because of having similar communication orientations and sharing similar knowledge of the localized application, as discussed in hypothesis 3.

2.3.2.2 Themes for sub-research question 2

The second sub-question focuses on the communication between the evaluators and users. This is a more explorative study compared to the study of the first sub-research question. Some communication patterns may exist in all cultural settings, whereas others may be only for the evaluators or users with specific cultural backgrounds. Since this is a more explorative study, this researcher did not have very clear hypotheses formed beforehand. Instead, some themes were developed based on the theories. From investigating the themes, the researcher hoped to get a clearer view of the evaluators' and users' communications in usability testing; thus, all the possible communication patterns are investigated in all cultural settings.

For the communication patterns, four themes are explored. The four themes are developed based on the consideration of the theories, previous work and the types of the codes. Theme 1, theme 2 and theme 3 concern the communication patterns in usability testing. Theme 4 is regarding the communication patterns within the usability problem happening period. Since the purpose of usability testing is to find usability problems, through theme 4, we could see the evaluators' and users' communication patterns when they were communicating the problems. The four themes are:

Theme 1: The time spent on the users' three state events---thinking aloud, talking rather than thinking aloud, and silence --- is different in different cultural settings.

As discussed in section 2.3.1.2, the three state events were put forward by considering Ericsson and Simon's thinking aloud model (1993), and Boren and Ramey's thinking aloud model (2000). With this theme, we can see which model is more suitable for usability testing and whether there is any difference in different cultural settings. From the culture theories, we can assume that Danish and Chinese users may spend different time on the three state events. For example, the time on thinking aloud for Danish users may be longer than that for Chinese users, whereas the time on silence for Chinese users may be longer than that for Danish users. According to the discussion in section 2.2.2.2, Chinese people, who tend to be holistic thinking, may not be as good at thinking aloud as are Danish people, who tend to be analytical thinking. Further, Chinese users who have high-contextual communication may not speak everything in their mind since they think some information is known by the evaluator, whereas Danish users who have low-context communication may tend to speak everything in their mind (section 2.2.3.3). The silent period may thus be longer for the Chinese users than it is for Danish users, and the thinking aloud period may be longer for Danish users than it is for Chinese users.

Theme 2: Users' verbal behaviors are different in different cultural settings.

Theme 2 is explored through the codes of the users' point events. Users' verbal behaviors are coded as negative comments, positive comments, suggestions and culture related comments (section 2.3.1.2). Because of the task-focus and socio-emotional focus orientations discussed in section 2.2.2.2, the Danish users may give more negative comments than do Chinese users, whereas the Chinese users may give more positive comments than do Danish users. Moreover, the task-focus and socio-emotional focus orientations also imply that when with foreign evaluators, the Chinese users' behavior may be influenced more than that of Danish users (section 2.2.2.2), which means that Chinese users behave more differently to local and foreign evaluators than do Danish users.

Theme 3: Evaluators' verbal behaviors are different in different cultural settings.

Theme 3 is investigated through the codes of the evaluators' point events. From theme 3, we explore whether Danish and Chinese evaluators behave differently and whether evaluators behave differently to the users from their culture, and the users from another culture. Because of the different communication orientations (low-context vs. high-context communication in section 2.2.3.3) and different cognitive styles (analytic vs. holistic in section 2.2.2.2), Danish and Chinese evaluators may behave differently in order to find usability problems. For example, Danish evaluators may tend to probe more questions than do Chinese evaluators, since they have low-context communication and they need to get more information in order to understand. Moreover, because of the users' difference, in order to get the users' thoughts, the evaluators may behave differently with Danish and Chinese users. For example, since Chinese users may not be good at thinking aloud, evaluators may give more reminders to Chinese users than to Danish users.

Theme 4: The way to communicate usability problems is different in different cultural settings.

Theme 4 is about the evaluators' and users' communication patterns for the usability problems. In order to better understand the way to communicate the problems, UPU and UPE are put forward. UPU (the Usability Problem is learnt from the Users) means that the usability problem information is transmitted from users to evaluators, whereas UPE (the Usability Problem is learnt from the Evaluator) means that the usability problem information is transmitted from evaluators to users. UPU is normally found through the observation of the user's task performing and verbal behavior. UPE is normally found through the evaluator's directed probing behavior. The directed probing means that the evaluator calls the user's attention to the issue that is not related to the user's current behavior. The coder coded UPU and UPE based on the observation of the video (see appendix 14). The division of the UPU and UPE is whether the usability problem is from the users' behavior or from the evaluators' mind. Both Ericsson and Simon's thinking aloud model (1993) and Boren and Ramey's thinking aloud model (2000) emphasize the users' role in finding problems, such as the evaluator's role of a passive listener in Ericsson and Simon's model, and "user-expert and evaluator-learner" in Boren and Ramey's thinking aloud model (section 2.1.3.1 and section 2.1.3.2). However, in the usability practice, evaluators are also actors and play the role of finding usability problems (Shi, 2008b). Previous study shows that evaluators ask questions about nonexistent parts of the system, speculative or hypothetical questions (Nørgaard & Hornbæk, 2006), which implies that the evaluator's own mind may call the user's attention or guide the user's direction. The field study in this project also shows (Clemmensen & Shi, 2008; Clemmensen et al., 2007; Shi, 2008a) that the evaluators in the industrial area liked to use probing to direct the users' attention to a specific issue that the user did not notice or did not mention. Using the communication process model to discuss (section 2.2.3.1), for the UPU, the user is the sender and the evaluator is the receiver, whereas for the UPE, the evaluator is the sender and the user is the receiver in identifying the usability problems. The communication patterns on UPU and UPE may not be the same, and thus it is necessary to divide the usability problems into UPU and UPE. In order to ensure that the evaluators' and users' communication patterns are different in finding UPU and UPE, the sub-themes about UPU and UPE are developed.

Theme 4a: The number and duration of the UPU and UPE may be different in different cultural settings.

Theme 4a is the investigation of the UPU and UPE in different cultural settings. The number and duration of UPU and UPE may be different in different cultural settings. For example, according to the "perception of control" in section 2.2.2.2, Danish evaluators may be more active than are Chinese evaluators, so the opportunity for Danish evaluators to start a usability problem may be greater than that for Chinese evaluators to start a usability problem. Therefore, Danish evaluators may tend to have more UPE than Chinese evaluators.

Moreover, since the problem information in UPU is transmitted from the user to the evaluator, and the problem information in UPE is transmitted from the evaluator to the user, the time on discussing the UPU and UPE may be also different in different cultural settings. For example, the time on UPU for distant pairs may be longer than that for local pairs. Since local pairs share the same cognitive styles and communication orientations (section 2.2.2.2 and section 2.2.3.3), it may be easier for local evaluators to understand the users' problems, which may result in less time spent on discussing usability problems, compared to the distant pairs.

Theme 4b: The time spent on discussing UPU/ UPE with different severity is different in the four cultural settings.

Theme 4b is used to examine whether evaluators and users spend more time on discussing the critical or important usability problems than on the minor problems. Error handling time is an important objective measure of the severity (section 2.3.1.1). In this study, the relation between the time on the problem and the problem severity is examined to see whether evaluators estimate the severity of the usability problem based on the problem handling time. Hassenzahl (2000) found a fundamental lack of correspondence between the objective and subjective measurement.

However, he did not consider the cultural issues and the way of finding the problems. UPE is from the evaluators' pre-thoughts, so there may be no clear relation between the time on the problem and the problem severity. UPU is the one learnt from the users, and the evaluators, especially foreign evaluators, may decide on the severity of the usability problem by considering the objective measure—time. For foreign evaluators, they get the UPU from the target user groups and they may not be familiar with the usability problem in the target culture. Thus, compared to local evaluators, foreign evaluators may rely more on the objective measures to estimate the severity of the problems, which may show a clear relation between the time on the usability problem and the severity of the problem.

Theme 4c: Users' communication patterns for the UPU and UPE are different in different cultural settings.

Theme 4d: Evaluators' communication patterns for the UPU and UPE are different in different cultural settings.

Theme 4c and theme 4d are developed based on the same reasons as theme 1 to theme 3. Users' and evaluators' communication patterns vary in different cultural settings in the usability testing. The communication patterns may also be different in discussing the usability problems because of the different cognitive differences and communication orientations (section 2.2.2.2 and section 2.2.3.3). Moreover, UPU and UPE involve different information transmission, thus evaluators' and users' communication patterns may be different for finding different types of problems.

2.3.3 Summary

Section 2.3 has discussed the two sub-research questions and then proposed hypotheses and themes based on the theories. The first sub-research question is about the evaluators' usability problem finding and usability problem severity rating behaviors in different cultural settings. Four hypotheses have been put forward in order to examine the sub-research question 1. The second sub-research question concerns evaluators' and users' communication patterns. Sub-research question 2 is a more explorative study; accordingly, four themes instead of hypotheses are proposed.

2.4 Chapter Summary

In this chapter, thinking aloud models, culture theories and the research question have been discussed. Based on the theories, hypotheses and themes are put forward. For the thinking aloud models, Ericsson and Simon's model (1998; 1993) and Boren and Ramey's model (2000) are introduced. In this study, the latter one is the main focus since it is proposed for usability tests, whereas the former one is proposed based on the information processing theory for investigating cognitive processes. For culture theories, Nisbett and his colleagues' studies of cognitive styles and Hall's high- and low-context communications are the main concern. The two theories are not separate, but connected to each other. Danish people are regarded as having analytic cognitive style and low-contextual communication orientation, whereas Chinese people are regarded as having holistic cognitive style and high-contextual communication orientation. According to the theories, hypotheses are developed for sub-research question 1, and themes are proposed for sub-research question 2.

In the following chapter, the methodology is discussed.

3 Methodology

This researcher holds a realism view on science which assumes that phenomena and processes exist outside our experience, and it is possible to obtain the information about the world through appropriate methods (Bryman, 2008). The research question in this dissertation is: "To what extent does the evaluators' and users' cultural background influence the thinking aloud usability testing?" The research question indicates a causal link between cultural backgrounds and the thinking aloud usability testing, so it is appropriate to conduct an experimental study to investigate this question (Creswell, 2003; Field & Hole, 2003; Maxwell & Delaney, 2004). Figure 11 shows the methodological framework for this research.

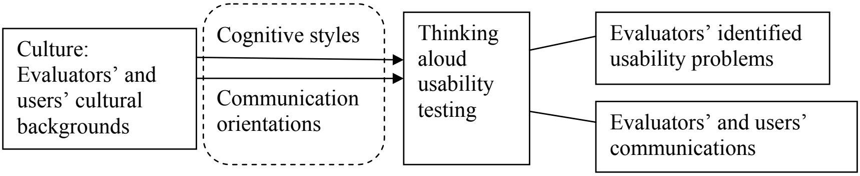

Culture: Evaluators' and users' cultural backgrounds
Cognitive styles
Communication orientations
Thinking aloud usability testing
Evaluators' identified usability problems
Evaluators' and users' communications

Figure 11: Methodological Framework

The two culture theories in the research are integrated in the conceptual framework. Evaluators and users with different cultural backgrounds may have different cognitive styles (Hall, 1989a; Hall & Hall, 1990) and communication orientations (Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002a), according to the theories discussed in chapter 2. From the culture theories, we have seen that culture may influence the thinking aloud usability testing, but the extent to which it does, and how it influences the testing are not so clear; accordingly, this empirical study was conducted to investigate the cultural influences. As discussed in chapter 1, this research, focuses only on examining the cultural effect on the identified usability problems and the communications in the thinking aloud usability testing.

Culture schemas are the characteristics of the participants, and are not given by the researchers. Vatrapu (2007, p.100) purports that "cultural schemas and cultural models are what human beings bring to an interactional situation…Therefore they are not rigid causal determinants of human behaviour. However, they can be strong influencers of human behaviour in appropriate situations and acceptable contexts." According to Vatrapu's study (2007, p.100), the relation between the culture and the thinking aloud usability testing (the arrow in the framework) is not "causal determinants," but "causal influencers." The experiment is conducted to investigate the influences of culture on the thinking aloud usability testing.

An experiment is usually done with a controlled research design which may change the context of that which is being investigated in the real situation and reduce the external validity (Maxwell & Delaney, 2004; Punch, 2005). However, the thinking aloud method itself is usually associated with laboratory testing (Norman & Panizzi, 2006); and the usability testing is also called a "laboratory usability testing" (Hartson et al., 2001, p. 374), which means that the form of the thinking aloud usability testing is the same as that of the lab experiment. Even though we perform the experiment to investigate the thinking aloud usability testing, it will not give more artificial or unnatural control than the test itself. The lab environment in this study is prepared exactly as the usability testing room in practice; thus, the evaluators are usability professionals and the users are potential real users of the application. Carrying out the experiment does not reduce the external validity so much in this research. However, since the applications, tasks and testing procedures are prepared by the researcher, the external validity may be affected to some extent (see section 6.3 about the external validity of this research).

In this chapter, the research design is presented first, after which the participants, materials, procedures of the experiment and data analysis are introduced.

3.1 Method

3.1.1 Research Design

In order to figure out the effect of culture on the thinking aloud usability test, an experimental study with a 2x2 factorial design (Cozby, 2003) was conducted in Denmark and China. The two independent variables are: evaluators' cultural backgrounds and users' cultural backgrounds. Each independent variable has two levels: Danish and Chinese (see Table 2).

Table 2: 2x2 Factorial design

	Users' cultural background	
Evaluators' cultural background	Danish	Chinese
Danish	Danish evaluators-Danish users	Danish evaluators-Chinese users
Chinese	Chinese evaluators-Danish users	Chinese evaluators-Chinese users

Table 2 shows the experimental design in this study. Since Danish people are regarded as having an analytic cognitive style and low-contextual communication, and Chinese people as having a holistic cognitive style and high-contextual communication, the experiment was conducted in both Denmark and China with both Danish and Chinese evaluators and the target users.

In this 2x2 design, there are four conditions labelled as four cultural settings: Danish evaluator-Danish user pairs, Chinese evaluator-Danish user pairs, Chinese evaluator-Chinese user pairs and Danish evaluator-Chinese user pairs. The four cultural settings can be grouped into local and distant settings. The local settings refer to the Danish evaluator-Danish user pairs and Chinese evaluator-Chinese user pairs, and the distant settings refer to the Chinese evaluator-Danish user pairs and Danish evaluator-Chinese user pairs. Moreover, the four cultural settings can also be grouped into Danish and Chinese settings. Danish settings are the tests in Denmark with Danish users and with both Danish and Chinese evaluators; Chinese settings are the tests in China with Chinese users and with both Chinese and Danish evaluators. Table 3 shows the complete research design of this study.

Table 3: Research design

Cultural profile pairs	Before the test	Usability testing	After the test
In Denmark: Danish evaluators-Danish users	Consent form, Background questionnaire	The tests conducted by the evaluators	Usability problem list for the users, Interview the users, Usability problem form for the evaluators after each test, Interview the evaluators after finishing the four tests, Evaluators' and users' communication behaviors coded by two researchers
In Denmark: Chinese evaluators-Danish users			
In China: Chinese evaluators-Chinese users			
In China: Danish evaluators-Chinese users[1]			

[1] There were two tests with one Danish evaluator and two Chinese users conducted in Denmark, since two of the four tests in China were failed and the Danish evaluator did two more tests in Denmark. The two Chinese users who attended the tests in Denmark were born and grew up in China. One of them had been in Denmark for less than one year and the other had been in Denmark for around two years. According to Clapp (1993, cited from Bull & McCalla, 2002), the cognitive style can be quite stable over three and a half years, so the two tests in Denmark will not influence the result of the study.

In order to avoid sampling bias, more than one evaluator was used in each condition. Accordingly, 16 evaluators attended this study with each evaluator doing four tests, 64 tests totally. In each condition, there were 4 evaluators with 16 tests. Each test lasted about 2 hours, about1 hour for the usability test, and 1 hour for the questionnaire and interview.

The tests in each condition should be comparable, and some issues should be controlled in order not to influence the result. These controlled issues are called control variables. Control variables (Punch, 2005) are extraneous factors which may influence the experiment and should be constant so as to minimize the effects on the result. In this research, the control variables were:

- Experience of being the usability professional
- Spoken English skills
- Familiarity with the application being tested

The above factors were similar in each condition, and the data were observed from the background questionnaire.

According to the two sub-research questions, there were two main dependent variables: the usability problems and the communication patterns. The usability problems were measured as:

1) Usability problem discovery: to examine whether local and foreign evaluators found similar usability problems. Usability problems in this study were explicitly clarified to evaluators in order to ensure that they had the same understanding of the application's usability and to minimize individual differences. The evaluators were told that usability in this study included both the products' attributes and the users' experience, and usability problems were those factors that made the system difficult to operate or an indication that they were not satisfied with the users (section 2.1.1.1 and section 2.1.2).
2) Unique and shared usability problems found by local and foreign evaluators: to examine whether there was a tendency for finding some specific types of usability problems by local and distant pairs (section 2.3.1.1).
3) Severity of the usability problems-- minor, important and critical: to examine the problem severity rated by Danish and Chinese evaluators (section 1.2 and section 2.3.1.1).
4) Users' agreement for the usability problems found and for the problem severity rated by the evaluators: the users were asked to come up with the usability problems and rate the severity of those problems that they had experienced in the test. Hartson et al. (2001, p. 387) postulate that "usability is ultimately determined by the end user, not an expert

evaluator, so realness of problems needs to be established by the user." Hornbæk (2009) argues that usability problems found by the thinking aloud usability testing may not be as real as researchers have thought. This study analyzes whether the users agree with the usability problems found and problem severity rated by the evaluators. Compared to the foreign evaluators' usability problem findings and severity ratings, local evaluators' identified problems and problem severities may be agreed upon more often by target users.

The first two measurements were used to examine the local and foreign evaluators' usability problem findings. The third was used to examine the severity ranks given by the evaluators with different cultural backgrounds. By examining the first three, we could see whether there were differences between the local and foreign evaluators' identified problems, which related to Hypothesis 1 and Hypothesis 2. The fourth, used to examine the users' agreement on the evaluators' usability problem findings and problem severity ratings, related to Hypothesis 3 and Hypothesis 4.

Theme 1 to theme 4 concerned the evaluators' and users' communication patterns. Communication patterns were measured by a well-defined coding system (section 3.5.3.2). Content analysis, a technique to analyze all kinds of verbal, pictorial, symbolic, and communication data (Krippendorff, 2004), was used to analyze the communication. One of the best known definitions of content analysis is given by Berelson (cited from Bryman & Bell, 2007, p. 302): "Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication." Content analysis can be used for both quantitative and qualitative studies (Krippendorff, 2004; Weber, 1990). In this research, quantitative analysis was mainly used to compare whether there was any significant difference between the communications with Danish and Chinese people. The communication contents were coded, based on the thinking aloud models and previous studies, which is introduced in section 2.3.1.2 and section 3.5.3.

3.1.2 Instruments

3.1.2.1 Consent Form and Background Questionnaire

In this study, the consent form and the background questionnaire were required to be filled out. The consent form showed that the participants were willing to be video and audio recorded for our research (see Appendix 1). The background questionnaire included two parts (see Appendix 2 and Appendix 3): 1) Part 1: the participants' demographic information; 2) Part 2: the participants' cultural orientation. Part 1 was used to collect the evaluators' and users' demographic information, such as age, gender, nationality, etc., as well as their level of English skills and the familiarity of the prototype in the test. Part 2 was used to examine the participants' cultural orientations[2]. There were two questions in Part 2, which have been used by Nisbett and his colleagues to examine people's ways of categorizing objects---based on relation, family resemblance or based on rules (Nisbett & Masuda, 2003). Of course, two questions alone cannot examine the participants' cognitive styles and communication orientations. But since the tests were conducted in the target country and all the local users and evaluators were born and raised in that country, and all the foreign evaluators were born and raised in the other country, coming to the target country mainly for this research (see section 3.2.3), the participants could represent the Danish and Chinese people described by the theories on some level.

In this dissertation, the participants' characteristics were based mainly on the existing culture theories and their findings (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002b). The focus of this research is the cultural variation in approaches to the thinking aloud usability testing. The studies conducted by Nisbett and his colleagues have shown the different cultural orientations in the West and East Asia, and thus making it unnecessary for this researcher to examine the participants' cultural orientations. Only two questions were used as detectors to see whether Danish and Chinese participants were different in this research. In the future, it would be better to develop some more reliable questionnaires or tests to examine the participants' cognitive styles and communication orientations.

3.1.2.2 Usability Problem List for the Users

After each test, the users were required to mark the problems on the usability problem list that they had mentioned in the test and also to give the severity of the problem. The usability problem list was developed from the pilot study (see Appendix 4). Section 2.3.2.1 and section 3.1.1

[2] Part 2 was not in the original background questionnaire, but in a separate cultural orientation questionnaire which was developed by this researcher without checking the validity. So in this PhD thesis, in order to make sure the questions used in this study were valid, only the two questions used in Nisbett and his colleagues' studies were analyzed.

indicated that the consistency of the usability problems found by evaluators and users would be analyzed in this study, to see whether the usability problems found by the evaluators were problems for the users and whether the evaluators and users gave similar severity ranks to those problems. However, since the users were not usability professionals and they may not have been clear about what usability problems were, it may have been difficult for them to write down the usability problems after the test. Nielsen (1992) claims that usability specialists found two to three times as many usability problems as non-specialists. Thus, in order to examine whether the problems found by evaluators were problems for the users, it was deemed a better solution to show users all potential problems and ask them to mark those that had occurred in the test. For this purpose, a usability problem list containing most potential problems was given to the users. If a usability problem was found by the evaluator, and this problem was also evident to the user, but the user did not mark it, then it may not have been a problem for this user, suggesting that there was disagreement between the evaluators and users. On the other hand, if a usability problem was marked by the user, but was not written down by the evaluator, it does not mean that this problem was missed by the evaluator, since this problem may not have been discussed or did not occur in the test. The user may have thought it was a problem when he/she read it on the list, but it may not have been mentioned in the test. Hence, in this research, we could only investigate the usability problems written down by evaluators, and whether the users agreed or not. We did not investigate whether the evaluators found the problems marked by users. Sub-research question 1 indicated that this study focused only on the evaluators' usability problem findings and severity ratings; thus, it is appropriate to focus the investigation mainly on the evaluators.

The usability problem list in this study came from the pilot study (Clemmensen, 2006; Clemmensen & Goyal, 2005; Sun & Shi, 2007), for which the prototype was similar to the one in this research. In the pilot study, there were 9 tests in Denmark and 6 tests in China. The usability problems were extracted by two researchers: one Dane and one Chinese. The researchers watched the videos of the usability tests in their own country and came up with two lists of problems. The two lists of usability problems were integrated into a new list of non-overlapping problems. There were 35 non-overlapping usability problems with a section for "others" to encourage the users to write down the problems that were not included in the list.

3.1.2.3 Usability Problem Form for the Evaluators

After the test, the evaluators were given a form to write down the usability problems that they had found in the test and rate the severity of each problem (see Appendix 5).

In this study, the rating scale for the severity of the usability problems was:

- M= minor usability problem- fixing this when an opportunity arises
- I= important usability problem- important to fix, so it should be given high priority
- C= critical usability problem- imperative to fix this before the system is put into use

3.1.2.4 Interviews for the Evaluators and Users

After the tests, interviews were conducted to get more information from the users and evaluators. From the interviews with the users, we got information about the users' experience of using English to think aloud and their feeling of doing tests with local and foreign evaluators (see Appendix 6). After all four tests were done by each evaluator, the researcher interviewed the evaluators about their criteria for choosing the usability problems, that is, deciding on the severity of the problems and their experience of doing the tests with local and foreign users (see Appendix 7).

3.2 Participants

3.2.1 Danish and Chinese Participants

In this research, Danish and Chinese evaluators and users were chosen to be investigated. The reasons of doing tests with Danish and Chinese participants were:

- First, this research was supported by the Danish government, so the findings were expected to be useful for usability studies in Denmark.
- Second, this PhD project focused on investigating the usability testing in Denmark and China.
- Third and most important, Danish and Chinese cultures represented the cultures discussed in the culture theories. As mentioned earlier, according to Nisbett and his colleagues, Denmark belongs to Western countries and China belongs to East Asian countries. Danish people tend to have an analytic cognitive style and low-context communication orientation, whereas Chinese people tend to have a holistic cognitive style and high-

context communication orientation (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Norenzayan, 2002a).

In general, even though it seems very simple to distinguish the world into the West the East Asia, and distinguish people in different cultures into analytic vs. holistic cognitive styles, and high-context vs. low-context communication orientations, the division does show that Danish and Chinese people are different in many aspects. The cultural dimension model does not indicate the "dichotomy," but "relatively ordered continuum" (Vatrapu, 2007, P. 151). Not all Danish people have an analytic cognitive style and low-context communication orientation, and not all the Chinese people have a holistic cognitive style and high-context communication orientation. However, previous studies have shown that Danish and Chinese people tend to have different cognitive styles and communication orientations (Hall, 1990; Hall & Hall, 1990; Nisbett & Masuda, 2003; Nisbett & Norenzayan, 2002a). This study investigates the Danish and Chinese evaluators' and users' communications, and the evaluators' usability problem finding and severity rating behaviors at the group level.

3.2.2 Language

In this research, the tests done by local evaluators and foreign evaluators were compared. Since all foreign evaluators did not speak the local language in the target country, in order to make sure the difference was not because of the language but other deeper factors, all participants spoke English.

It was not difficult to select Danish people who could speak English well, because there are many courses available in English universities and many people speak English often in their work. It could have been an issue for Chinese people; however, all Chinese evaluators in this study spoke good English. We interviewed the Chinese users in English to test their English skill before they came to the test. All Chinese participants needed to have passed the TOEFL (Test of English as a Foreign Language, above 580 points or 90 points) or IELTS (International English Language Testing System, above 6.5 points), or College English Test Band 6, or PETS5 (Public English Test System) or TEM 4 (Test for English Majors) and to be good at spoken English.

There could be selection bias when using users who can speak English well. But it is not a serious issue compared to the influence of asking local pairs and distant pairs to speak different languages (Sun & Shi, 2007). In order to make sure that the differences we found in this study were deeper cultural issues, such as cognition styles and communication orientations, it was better to ask all participants to speak English. Thus, the language was controlled in all conditions.

Further, since this was an exploratory formative evaluation, the purpose of which was not to see the efficiency of doing the tasks but to find usability problems, as long as the users could clearly express their thoughts, the language used in the test should not influence the results as much as in summative evaluation. Researchers have found that "even though language of testing may lead to changes in bilinguals' responses, such changes occur in a limited range and do not necessarily threaten the conclusion pertaining to cultural comparisons" (Ji, Zhang, & Nisbett, 2004, pp. 64-65). This indicates that since all participants spoke English, it should not influence the comparisons of the tests with participants from different cultures.

3.2.3 Participant Selection

All evaluators in this study were experienced usability practitioners who had previously done usability tests. The Danish/ Chinese evaluators and users were born and raised in Denmark/ China. They could speak English fluently and were familiar with Microsoft Word.

The evaluators were recruited not only through emails to usability specialists that who introduced by colleagues, but also by putting recruiting information as a part of the newsletter in the SIGCHI (Special Interest Group on Computer-Human Interaction) Denmark. The travel fees for foreign evaluators were paid by the project and a gift card was given to local evaluators for their effort in the study. The Danish users were recruited mainly through distributing emails to the students in CBS, putting information in the campus bulletin boards, and asking a usability consulting company (Snitker & Co.) to find participants from their user network (Appendix 8 shows an example of the email for recruiting Danish users). The Chinese users were recruited mainly through putting recruiting information on campus bulletin boards.

3.3 Materials

This section introduces the application used in this research, the tasks in the tests, and the usability lab and equipments.

3.3.1 Application

In this research, a culturally localized prototype was used as the testing subject in the thinking aloud usability tests. In this section, the concepts of localized application and prototype are clarified first, and then the prototype used in this research is introduced.

3.3.1.1 Localized Application vs. Globalized Application

As introduced in section 2.1.2, the application could be developed in two ways: localization and globalization. The aim of globalization is to create application/ products with universal design. The universal designed application is local neutral and usable for all people without the need for adaptation or specialized design (Horton, 2005). Since people are very different in various cultures, it is almost impossible to design an application which could be used by people all over the world. In order to ensure that an application can be used by people from different cultures, applications need to be localized for different target user groups. Localization seeks to "create custom versions for each locale" (Horton, 2005, p. 158). Users in different locales are fundamentally different and require a highly customized user interface. How a person interprets the interface is affected by the person's cultural background (Horton, 2005). For example, when seeing images with a lot of red color, Danish people may think of Christmas, but Chinese people may think of weddings. Because of their customs, they are "much more likely to recognize images of common objects that resemble the ones they see every day" (Horton, 2005, p. 159).

The culturally localized application may influence the thinking aloud usability testing which is facilitated by foreign evaluators. In Vatrapu and Pérez-Quiñones's study (2006), the website being tested was a culturally localized website, which meant that people in other cultures might not understand the background, purpose and other detailed issues of it. On the one hand, it was not easy for a foreign interviewer to find the culturally sensitive usability problems. On the other hand, the users of the study did not talk very much with the foreign interviewer since they thought the foreign interviewer did not understand it. The users with the foreign interviewer just gave their opinions with little communication and interaction with the interviewer which, in turn, influenced the usability problems that the foreign interviewer would find. This implies that compared to the international application, the culturally localized application is more sensitive to the interviewers' and users' cultural backgrounds.

In the current study, a culturally localized application was used, not a globalized application, since: 1) A culturally localized application could act as a primer to elicit a user's culture related communication or behavior, and 2) Usually the purpose of doing usability testing in different cultures is to see whether this application is accepted by users in the target culture, which is the localization process. Moreover, in order to make sure that the applications were comparable across cultural settings, in this study, localized versions of the same application, rather than indigenous applications, were used.

3.3.1.2 Prototype vs. Final Application

Section 2.1.2 introduced two kinds of applications: prototype and final application. The prototype is an unfinished application which can reflect the final application in some way. In formative usability testing, the subject being tested is usually a prototype (Barnum, 2002; Nielsen, 1993). The prototype, which is cheaper and can be developed faster, is the main test subject in the early usability evaluation (Nielsen, 1993, p. 93). A prototype can be produced by (Nielsen, 1993, pp. 95-97):

- Placing less emphasis on the efficiency of the implementation
- Accepting less reliable or poorer quality code
- Using low-fidelity media that are not as elaborate as the final interface but still represent the essential nature of the interaction
- Using fake data and other content
- Using paper mock-ups instead of a running computer system

Some of the methods proposed by Nielsen were extracted to produce the prototype. The entire idea behind testing a prototype is to "save on the time and cost to develop something that can be tested with real users" (Nielsen, 1993, p. 94). If the application is already finished, it will be very hard to change. Instead, testing prototypes with different fidelity levels can get users' feedback through the whole application development process, which will produce satisfied application for the users in the end; thus the important or main job for doing formative evaluation with prototypes is to get as much diagnostic information as possible, which implies that skilled communication with users is very necessary. Since thinking aloud usability testing is normally used as formative evaluation, it was appropriate to use an unfinished prototype as the testing application in this study.

3.3.1.3 The Application being tested in this Research

The application in this research was a “wedding invitation application” prototype. The prototype was designed by adding a collection of wedding images and icons to My Collections in Microsoft Word’s Clip Art organizer (Clemmensen & Goyal, 2005; Sun & Shi, 2007). Clip Art in Microsoft Word was used in the pilot study (Clemmensen, 2006; Clemmensen & Goyal, 2005; Sun & Shi, 2007). The prototype could be regarded as culturally localized, since Microsoft Word had different versions in Denmark and China; the focus was also Cultural Clip Art, not Microsoft Word. In order to make the tasks easier for the users, we also provided some wedding invitation texts and backgrounds.

This application has been used for people in different cultures to design wedding invitations. Chinese and Danish wedding invitation prototypes were the sub-applications. Each sub-application included three parts, images (in the Clip Art organizer), texts and backgrounds, see Figure 12. In the usability test, the usability of each sub-application was evaluated in the corresponding country.

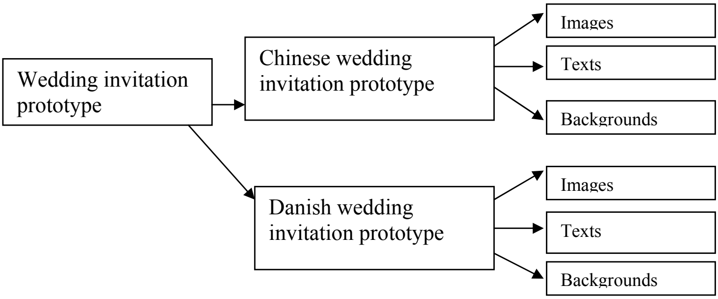

Wedding invitation prototype
Chinese wedding invitation prototype
Images
Texts
Backgrounds
Danish wedding invitation prototype
Images
Texts
Backgrounds

Figure 12: The application being tested in the study

The prototype was developed separately for Chinese culture and Danish culture, but considered similar rules. The images were selected from the image folder in the pilot study (Clemmensen & Goyal, 2005; Clemmensen, Yammiyavar, & Sun, 2006). In the pilot study, the images were collected through the Internet. Questions about images of the wedding invitation were put on some well known Bulletin Board Systems. By considering the suggestions of the images given by people online, around 100 images were selected by Chinese researchers for the

Chinese wedding invitation prototype, and around 100 images were selected by Danish researchers for the Danish prototype. In this research, considering the time consuming activity of selecting so many images, 50 images were selected from the image folders in the pilot study for the Chinese prototype, and 50 images were selected for the Danish prototype. Cultural issues were introduced to the image folder in order to see the differences between the tests done by local evaluators and foreign evaluators. The Chinese image folder of 50 images included 25 traditional Chinese wedding invitation images, 20 modern wedding invitation images, and 5 wrong images; the 50 Danish images also had 25 typical Danish wedding invitation images, 20 universal wedding invitation images and 5 wrong images. Since the Chinese modern wedding invitation was usually from Western countries and the images were universally as those used in the Western countries, such as rings, flowers, etc, the 20 modern wedding invitation images were the same as the 20 universal wedding invitation images in the Danish image folder in order to make the two prototypes comparable. Thus, there were 80 images picked from the two image folders in the pilot study consisted of: 20 universal wedding invitation images picked by the cooperation of one Chinese and one Danish researcher from all the 200 images in the pilot study, 25 traditional Chinese images were picked from the Chinese image folder in the pilot study by the Chinese researcher, 25 typical Danish wedding images were picked by the Danish researcher from the Danish wedding image folder in the pilot study, and 5 wrong images were selected for each cultural prototype in order to be used as potential usability problems (Vatrapu & Pérez-Quiñones, 2006). The five wrong images were those which would never be used for wedding invitations in the target culture, including images not related to weddings, or images related to wedding but not used in the target culture. Figure 13 shows the interfaces of the Danish and Chinese wedding invitation Clipart folders in the tests.

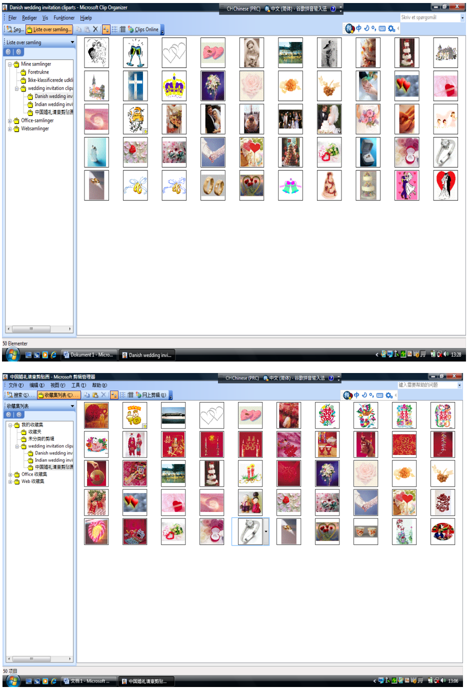


Figure 13: The Danish and Chinese wedding invitation Clipart folders

Since the texts were not easy to browse and users needed to read every one, in this study only 4 wedding invitation texts were provided in each cultural prototype. The texts were developed from the real wedding invitation texts given by colleagues of this researcher. All texts were in the local language.

Nine backgrounds were provided for each cultural prototype. Apart from one blank background, four of them were made from images, and the other four were made from the "Background" function in the "Format" menu.

3.3.2 The Task in the Test

All thinking aloud usability tests involve the selection of a set of target tasks for the users to perform during the testing session. It is impossible to test all the tasks that users will do in a real situation. There are two important criteria for selecting tasks: 1) selecting those that the users will do in the real situation, and 2) selecting those that could be diagnostic in revealing usability problems (Dumas & Redish, 1999; Norman & Panizzi, 2006). In this research, the tasks were designed by considering the purpose of the application and also the potential problems of the application.

The tasks in the experiment of this research were all necessary tasks that users would conduct with this application. In the tests, the goal of the prototype being tested was to make wedding invitations, and the general task was to make a wedding invitation that the users would want to use in their weddings. The task had been mentioned in the recruiting letters when recruiting the users (see Appendix 8). In the test, a short scenario was given to the user:

"Imagine you will get married and you want to make a wedding invitation by yourself. Please make a wedding invitation that you want to use in your wedding by using the wedding invitation application we provide."

The general task is divided into eight sub-tasks:

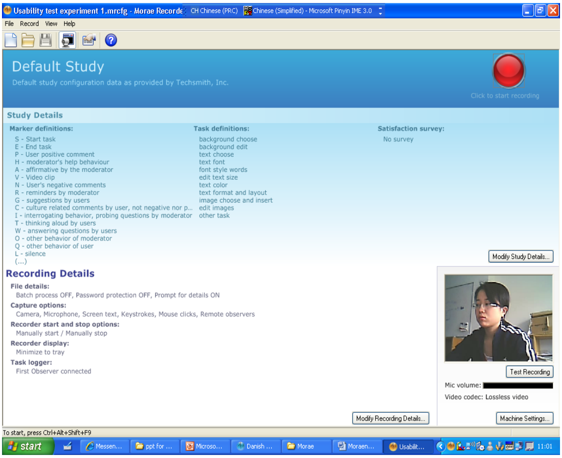
- You can choose a paper (background) for the wedding invitation if you want to.
- You can edit the background you chose if you think it is necessary.
- Please write the text that you want to appear on the invitation. (You can pick one from the text examples file, and you can also make some changes if you want to.)
- Please choose the appropriate font(s) for the text (such as the font, the font style, the size and so on).
- Please choose the color(s) for the text.
- You are free to choose any kind of formatting and layout for the text and make it prettier.
- Now choose one or some images from the Danish/Chinese wedding invitation Clipart sub-folder to make the wedding invitation look happy, colorful, and joyful, as this is for a wedding.
- Please edit the images if you want to (such as the size, brightness, contrast, layout, and so on).

Appendix 15/appendix 16 shows an example of a wedding invitation made by a Danish/Chinese user.

3.3.3 Usability Lab and Equipments

The usability laboratory usually has testing equipment and space for observation (Norman & Panizzi, 2006). In this research, since the experiments were conducted in both Denmark and China, it was better to use the same equipment in both countries to ensure that the equipment itself, such as the video camera, did not influence the results. (Portable equipment is a good way to do the cross-cultural tests in different countries.)

In this study, Morae was used to record the whole test process. Morae is a software-based solution for usability testing, which enhances data collection and speeds up analysis. It is comprised of three parts: recorder, observer, and manager. The three parts work together to provide a complete picture of the testing. With the Morae recorder (Figure 14), the screen and the keyboard activity of the user, the faces of the user and the evaluator (through a web camera), and the audio of the user and evaluator (through a microphone) can be recorded at the same time. Figure 14 shows the interface of the Morae recorder while the researcher is defining the tasks. The recorder runs silently in the background, and when it starts to work (pressing the red button) it will become a small icon on the right corner; however, people will not notice this and it does not disturb the user. With the Morae observer (Figure 15), the researcher can observe the screen and the video of the user and evaluator, hear the audio of the user's and evaluator's communication, log the tasks and take notes in another room. Figure 15 shows what the researcher can see through the Morae observer in the observation room. The Morae manager is usually used to analyze the behaviors that researchers are interested in. In this study, Morae manager was used to change the recorded videos to the form of ".wmv.", in order to ensure that the videos can be opened by most media players, such as Windows Media Player.



Default Study

Default study configuration data as provided by Techsmith, Inc.

Study Details

Marker definitions:

S - Start task
E - End task
P - User positive comment
H - moderator's help behaviour
A - affirmative by the moderator
V - Video clip
N - User's negative comments
R - reminders by moderator
G - suggestions by users
C - culture related comments by user, not negative nor p...
I - interrogating behavior, probing questions by moderator
T - thinking aloud by users
W - answering questions by users
O - other behavior of moderator
Q - other behavior of user
L - silence
(...)

Task definitions:

background choose
background edit
text choose
text font
font style words
edit text size
text color
text format and layout
image choose and insert
edit images
other task

Satisfaction survey:

No survey

Recording Details

File details:
Batch process OFF, Password protection OFF, Prompt for details ON

Capture options:
Camera, Microphone, Screen text, Keystrokes, Mouse clicks, Remote observers

Recorder start and stop options:
Manually start / Manually stop

Recorder display:
Minimize to tray

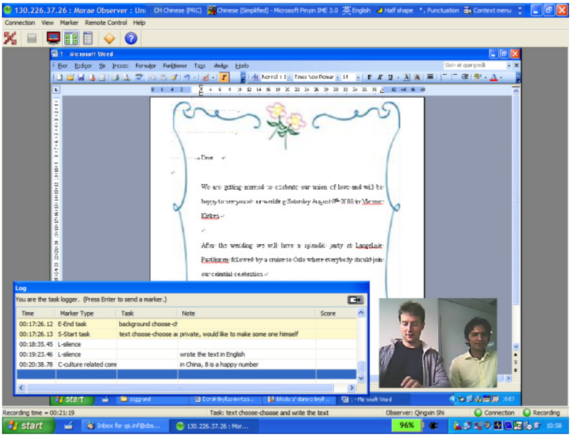
Task logger:
First Observer connected

Mic volume:

Video codec: Lossless video

To start, press Ctrl+Alt+Shift+F9

Figure 14: Morae recorder



Dear

We are getting married to celebrate our union of love and will be happy to see you at our wedding Saturday August 9th 2008 in Marmor Kirken

After the wedding we will have a splendid party at Langelinie Pavilionen followed by a cruise to Oslo where everybody should join our celestial celebration.

Log

You are the task logger. (Press Enter to send a marker.)

Time	Marker Type	Task	Note	Score
00:17:26.12	E-End task	background choose-cl		
00:17:26.13	S-Start task	text choose-choose ai	private, would like to make some one himself	
00:18:35.45	L-silence			
00:19:23.46	L-silence		wrote the text in English	
00:20:38.78	C-culture related comr		in China, 8 is a happy number	

Recording time = 00:21:19

Task: text choose-choose and write the text

Observer: Qingxin Shi

Connection

Recording

Figure 15: Morae observer

Two rooms were used in both Denmark and China: one testing room and one observation room. The equipment was the same in the two countries: a web camera, an audio recorder, and the same laptop with the prototype and the software Morae.

3.4 Procedure

Pilot studies with 9 tests in Denmark and 6 tests in China had been conducted before performing the formal experimental studies in Denmark and China. From the pilot study, the prototype, tasks, and protocols for the tests were modified for the formal experimental studies, and a usability problem list was developed for the experiment. As mentioned earlier, in the formal study, there were 64 tests totally. Each test was divided into three sessions: before the thinking aloud usability testing session, in the usability testing session and after the usability testing session. The usability testing session refers to the test done by the evaluators and users. Before and after the usability testing sessions were the sessions done by the researcher. The division of the three parts is different from that of the three parts shown in the research design (Table 3). The procedure of the formal experiment is introduced in greater detail in the following sections.

3.4.1 Before the Thinking Aloud Usability Testing Session

In the formal experiment, instructions and trainings were given to the evaluators before the tests (see Appendix 9). Even though all evaluators were usability professionals, it was still necessary to introduce the whole study in detail, since different tests had different purposes. The evaluators were informed about:

- The purpose of the test: to find usability problems and give severity of the problems after each test. The usability problems in this study included not only the problems of the functions, such as whether the application was easy to use, but also the content of the prototype, such as whether the users were satisfied with the images, backgrounds and texts provided. They needed to take notes of the usability problems during the test.
- The instructions to the users (see Appendix 10).
- The application in the test (see Appendix 10).
- The tasks in the test (see Appendix 11).
- The protocol of the test (see Appendix 12).
- The way to do the thinking aloud usability testing: they could communicate with the users during the test as they usually did.

We gave instructions to the evaluators that they could communicate with the users as they usually did, since how to do thinking aloud is not the same for every usability practitioners.

Whether the evaluator should communicate with the user during the thinking aloud test is arguable for researchers (Ericsson & Simon, 1993). But in the industrial area, most usability practitioners communicate with users when doing usability tests (Boren & Ramey, 2000). In order to ensure that the evaluators did the test in their normal way, it was felt that it was better to tell them they could communicate if they thought it was necessary for them to understand the user's speech in order to find the usability problems.

When the evaluators were familiar with the procedure of the test, they signed the consent forms and filled in the background questionnaire.

3.4.2 In the Thinking Aloud Usability Testing Session

The test started when the user arrived at the usability laboratory. The software Morae was used to record the sound, the screen and the faces of the evaluator and user through a web camera, and an audio recorder was used to record the sound as a backup. The researcher watched the test in the observation room during the test.

The users were given instructions of the test and asked by the evaluators to sign the consent forms (see Appendix 10 and Appendix 1). The demographic information was filled out by the users. The reasons for asking the evaluators, not the researcher, to give the users the consent forms and the background questionnaire were: 1) to enable the evaluators to have some time to prepare for the testing; 2) to help the users to relax (Dumas & Loring, 2008).

Before the thinking aloud usability test, the concept of thinking aloud was introduced to the users by the evaluators. In order to ensure that the users got the idea of thinking aloud, it was important to give examples of thinking aloud and have the users practice (Dumas & Loring, 2008, p. 67). In this study, apart from the introduction of thinking aloud, the evaluators also gave the users an example of thinking aloud by thinking aloud themselves while opening MS PowerPoint, and then asking the users to think aloud while they were opening Windows Media Player and selecting a song to listen as practice.

After introducing the application, the thinking aloud usability testing started.

The tasks were given to the users one by one. First, the general task description was shown to them, and then each sub-task was shown with the same order. When the users had finished all the tasks and the evaluators thought the test was over, the thinking aloud test session stopped.

During the test, the evaluators had to take notes which would help them retrieve the usability problems after the test.

3.4.3 After the Thinking Aloud Usability Testing Session

When the test was over, the researcher went to the test room and stopped the Morae and audio recording. The evaluator was given a usability problem form to write down the problems he/she found in the test and to indicate the severity of each problem (see Appendix 5). Since it was not necessary to talk with the evaluators while they were filling in the forms, the researcher was together with the users in another room.

The users were asked to mark the problems on the usability problem list that they had mentioned or that had occurred in the test, and to give the severity of the problems that they had marked. Afterwards, the researcher then interviewed the users to get more information about their experience and feelings in the tests. Before they left, the users were paid by the researcher.

After finishing the interview with the users, the researcher discussed with the evaluator to see whether he/ she had any questions while filling in the usability problem form or doing the test with the user. In other words, after all four tests were finished, the evaluator was then interviewed by the researcher about his/her criteria for deciding on the usability problems, on the severity of the problem and the cultural issues in the tests.

3.5 Data analysis

The two sub-questions in section 1.2 indicate that the research mainly involved the analysis of usability problems and the analysis of communications. In order to analyze the usability problems which were written down by the evaluators, and the evaluators' and users' communication patterns quantitatively, the data needed to be transformed. From the hypotheses in section 2.3.2.1, we know that the usability problem analysis involved two parts: comparing the usability problems found by the evaluators in different conditions, and comparing the problems found by the evaluators and users. The communication patterns were analyzed based on the thinking aloud models and culture theories (section 2.3.1.2). The evaluators' and users' verbal behaviors were coded according to a coding system introduced in section 3.5.3.2. In the following sections, we discuss how to analyze the usability problems and communications in this study.

3.5.1 Usability Problem Analysis: A Comparison of the Usability Problems Found and Rated by Evaluators in Different Cultural Settings

3.5.1.1 The Way to Match the Usability Problem Instances

The common way to analyze the usability problems is to produce a list of problems (Hartson et al., 2001; Skov & Stage, 2005). In this study, after each test, the evaluators were asked to indicate the usability problems that have occurred. The problems written down by the evaluators are called usability problem instances (Hvannberg et al., 2007). Even though usability problem counting is used as a main approach to compare different UEMs or different conditions of using a particular UEM (Hornbæk, 2009), it is criticized by some researchers (Gray & Salzman, 1998b; Hartson et al., 2001; Hornbæk, 2009; Keenan et al., 1999). Hartson, Andre et al.(2001) argue that the usability problem descriptions are given by the evaluators after the test, so the description may be expressed in whatever terms seem salient to the evaluator at that time (Hartson et al., 2001). Different descriptions may point to the same problem sometimes. Thus, it is necessary to develop a non-overlapping usability problem list. In the study conducted by Law and Hvanneberg (2004), the authors eliminated the duplicated usability problems and produced a non-overlapping problem list to analyze the identified problems in the tests conducted by users from different European countries. In this research, since some instances had similar meanings, the problem instances were organized into non-overlapping categories in order to compare the different problems found by the local and foreign evaluators. For example, a problem instance found with user 1 was "the images are not personal," and later this instance was found again with user 2 "personal images are needed." Although there were two problem instances, they were actually pointing to the same usability problem.

In order to decide which usability problems were different problems and which were overlapping problems, we needed to match the descriptions of the usability problems written down by the evaluators to see whether the descriptions had the same underlying design (Hornbaek & Frøkjær, 2008). Hornbæk (2009) warns that matching usability problem descriptions is not straightforward, but a difficult activity. Hornbaek and Frøkjær (2008) compare four techniques to match the usability problem descriptions: 1) the similarity of solutions to the problems, 2) a prioritization effort for the owner of the application tested, 3) a model proposed by Lavery, Cockton, & Atkinson (1997), and 4) the User Action Framework

(Andre, Hartson, Belz, & McCreary, 2001). They found that different techniques resulted in a different number of grouped usability problems and identified unique problems. In order to compare the usability problems in different cultural settings, the problem instances should be matched in the same way.

In this research, the problem instances were categorized into some non-overlapping usability problem types. Andre, Hartson et al. (2001, p. 110) stress that "classification of usability problems by type is not only valuable within the usability development process, but is also necessary for characterizing the strengths and weakness of usability evaluation methods within usability evaluation methods comparison studies." Since the description of the problem instances written by the evaluators after each test was brief and may not have included all the information (Hornbaek & Frøkjær, 2008; Lavery et al., 1997), it was better to organize the problem instances into types instead of many detailed usability problems.

The problem instances were thus categorized by considering: 1) the parts of the application, 2) the users' goals, and 3) cultural issues. First, we determined what the problem was about, whether it was about the images, text, backgrounds, tasks or other functions of MS Word. Second, for each part, we considered the users' intent of doing the task. For example, a problem instance was described as "do not know how to make the background of the images transparent" and another problem instance was described as "the background and images are not matched." For both problem instances, the users hoped to make the image and the background match. Even though the first description emphasized the function when doing the task and the second description emphasized the content of the application, the users' goals were the same---making them match. Accordingly, they were categorized as the same problem type. Since the researcher watched all the tests in the observation room, for most problem instances described by the evaluators, she was clear about the users' goals of doing the task. For unclear problem descriptions, the researcher asked the evaluator after the tests. Third, since culture was an important issue for developing the usability problem types in this study, we examined the problems found by evaluators within different cultural settings, as some culturally sensitive problems should not be grouped together with non-sensitive problems. For example, the aesthetic problem of the provided backgrounds was divided into two problem types: "inappropriate invitation background" and "background quality problems." "Inappropriate invitation background" refers to the cultural issues of the backgrounds, such as some backgrounds are not for Danish/Chinese wedding invitation. “Background quality problems” refers to the quality of the backgrounds and no cultural issues are involved, such as backgrounds are not pretty/ good enough.

3.5.1.2 The Way to Develop the New Usability Problem List in this Study

As we had developed a problem list in the pilot study (section 3.1.2.2 and Appendix 4), the new usability problem list in this research was developed by considering the original problem list from the pilot study. Both the original pilot study problem list and the new problem list in this study were developed as described in the previous section (3.5.1.1). The original problem list was developed by two researchers (section 3.1.2.2), whereas the new problem list was developed by the author of this dissertation by making some changes to the original list. The new problem list was not exactly the same as the original list, since:

1) The application in the experiment was different from the application in the pilot study. For example, in the pilot study we provided about 100 images in each sub-application. There was a problem called “image selection problems of too many images.” In the experiment, none of the evaluators found this problem, since there were only 50 images for each sub-application.
2) The test instruction was not the same as the one in the pilot study. In the pilot study, since there were no clear instructions, different evaluators did the tests in different ways, which resulted in a variety of problems. For example, in the pilot study in China, we did not open the clipart folders before the tests. The users had to find the clipart folder themselves, which caused some problems, such as “problem of the clipart search collection” and “keywords to find images are not clear.” In the experiment, the clipart folder opened before the tests, and thus these problems did not occur for the evaluators.
3) Some problems were not included in the original problem list and were added to the new list.

The new usability problem list was developed on the basis of the original list, but with some changes and according to the same categorization rules (section 3.5.1.1):

1) For the problems, which were not included in the original list, new usability problems were added to the new list.

2) The problems that were on the original list but not found in this study were deleted.
3) After adding and deleting some problems on the original list, all usability problems needed to be reorganized by using the categorization rules in order to make them much clearer: the parts of the application, the users' goals, and cultural issues.

3.5.1.3 Usability Problem Finding and Severity Rating Analysis

The usability problem finding analysis was based mainly on the problems on the new list. The detection rates, as well as shared and unique problems were analyzed in each cultural setting. The detection rate was defined as the average of the set of problems identified in each test divided by the set of problems identified collectively in all tests (Law & Hvanneberg, 2004): Detection rate p = Average of $|P_i| / |P_{all}|$.

In order to examine the Danish and Chinese evaluators' problem severity rating tendency, the severity ranks given to the problem instances were analyzed. It was difficult for the researcher to decide on the severity of the non-overlapping problems on the list, since each usability problem type may have had different problem instances with different severities. Different evaluators may "rate two instances of the same usability problem very different, depending on their effects upon the users" (Andre et al., 2001, p. 112). Thus in this research, the severity ranks given by the evaluators to the problem instances were analyzed after each test to examine their tendency of rating the problems, as described by Hypothesis 2.

Apart from comparing the evaluators' usability problems within different cultural settings, the problems found and rated by evaluators and users within each cultural setting were thus analyzed.

3.5.2 Usability Problem Analysis: A Comparison of the Usability Problems Found and Problems Severity Rated by Evaluators and Users in different Cultural Settings

3.5.2.1 Data Preparing for Analyzing the Evaluators' and Users' Usability Problems

As discussed in section 2.3.2.1 and section 3.1.1, "Usability is ultimately determined by the end user, not an expert evaluator, so realness of problems needs to be established by the user" (Hartson et al., 2001, p. 387). This research investigates whether the users agreed or not with the usability problems found and the problems severity rated by the evaluators.

The analysis of the comparison of the problems found by evaluators and users was based on the original problem list (Appendix 4) developed from the pilot study (section3.1.2.2). The users were required to mark the problems that they had mentioned or encountered in the test, since they were not usability professionals, it was hard for them to write down the usability problems (section3.1.2.3). The evaluators were also asked to write down the usability problems that they had identified during the test. The reason for not asking the evaluators to mark the problems on the list after each test was that we did not want to influence the evaluators' problem finding behavior. If the evaluators had been asked to mark the problems on the list as the users did, there may have been no difference between usability problems found by evaluators in different cultures, since the list could have made the evaluators mark the usability problems that they did not notice in the tests and thus would limit their thoughts of finding other usability problems. Of course, the list could have the same influence on the users' problem finding behavior. There were some problems that the users did not really mention in the test, but they still marked them, since they thought they were problems when they read the description of the list. However, since this study focuses only on analyzing the usability problems found by the evaluators, the research did not analyze whether the users agreed or not, nor did it analyze the usability problems marked by users but not found by the evaluators.

The evaluators' problem instances were classified into the corresponding usability problems on the list used by the users. If a problem was written down by the evaluator and also marked by the user, it meant that the evaluator and user agreed on that problem. If a problem was written down by the evaluator but not marked by the user, it indicated that the user did not think it was a problem, indicating a disagreement between the user and the evaluator. Since the users had seen all the problems on the list, they marked it if they thought they had the problem in the test. In this study, the problems which were not written down by the evaluator but marked by the user were not analyzed, mainly since we were not sure whether it was a real usability problem in this test.

The way to classify the evaluators' problem instances into the corresponding usability problems on the list is described in the following:

- Prepare 64 copies of usability problem lists for the 64 tests.
- For each test, read the problems written down by the evaluator and give each problem a number which corresponds to the usability problem number on the list. For the problem that does not belong to any number 1 to number 35 on the list, give number 36 "other" to it.

- The severity of the usability problem is the same as the one given by the evaluators. If two or more problem instances with different severities belong to the same usability problem on the list, the higher rank is given to that problem. For example, two problem instances written down by the evaluators belonged to the same usability problem number 27 but with different severities: 1) the backgrounds are blurry (important problem), and 2) the backgrounds are designed unprofessionally (minor problem). Therefore, the usability problem 27 is rated as being important.
- If description written down by the evaluators of one problem instance involves more usability problems on the list, then both problem numbers are given to the problem instance. For example, the evaluator wrote a usability problem as "all the background are not pretty enough, need to provide more backgrounds." This description involves both usability problem 27 and problem 30, and thus both numbers are given to this problem instance.
- Since the usability problem list was developed based on the pilot study, there were some problems that the evaluators identified in this study but for which they were unable to find the corresponding problem number on the list. So there were some "other" usability problems for the evaluators. Most "other" problems were not found by the users. However, this did not mean that the "other" problems brought forward by the evaluators were "false alarms," since the users were not usability professionals and it was reasonable for them to miss these problems (Nielsen, 1992). Since the "other" problems were beyond the users' ability to find, these problems were not included in this analysis.

The way to assign the problem instances written down by the evaluators on the usability problem list was to go through the description of the instances and give a corresponding problem number on the list. The main idea was to examine the understanding between the evaluator and the user. Considering the users' ability to find the usability problems, the analysis was based on the problem list which had been used by the users in the test, focusing on analyzing whether the usability problems detected by the evaluators were also considered as being problems for the users. The evaluators' "other" problems were not included, since they were not on the list and most users did not have the ability to detect them.

We conclude thus that the original usability problem list plays an important role in comparing the evaluators' and users' identified problems and problem severities. In each test, there were two usability problem lists, the evaluator's usability problem list and the user's usability problem list. When examining the users' agreement on the evaluators' identified usability problems, only the problems on the list are analyzed.

3.5.2.2 Analysis of the Severity Ranks

The analysis of the evaluators' and users' severity ranks was also based on the original usability problem list described in section 3.1.2.2. In order to examine whether the evaluator and the user give similar severity ranks, the usability problems should be found by both of them, since they are only required to rate the severity of the identified problems, that is, the data in the analysis is limited to the problems found by both evaluators and users.

The severity rating scale included three ranks: minor, important and critical. Since there were only three ranks, it was not suitable to use Pearson's correlation which is a parametric test and requires an interval scale of measurement. Spearman's correlation was used to calculate the correlation of the evaluators' and users' severity rating. Spearman's correlation is a non-parametric measure of correlation which works with ordinal-level data (Coolican, 2004, p. 443; Hinton, Brownlow, McMurray, & Cozens, 2004, p. 300). Spearman's correlation is used when there are no tied ranks. In this study, the ranks of the problem severity "minor," "important," "critical" were not tied. It was difficult to distinguish what was minor, what was important, and what was critical. The severity rating was mainly based on the users' and evaluators' own feelings. Accordingly, Spearman's correlation was used.

3.5.3 Communication Analysis

The basic idea of the communication analysis was to see the communication patterns when Danish and Chinese evaluators tried to find relevant usability problems with both Danish and Chinese users. The communication analysis included two parts: the analysis of the communication patterns in the test, and the analysis of the communication patterns for the usability problems. The purpose of the usability test was to find usability problems; apart from analyzing the evaluators' and users' communication patterns in the tests, it also analyzed the communication patterns in the usability problem discussing period. In the following sections, the video selection, coding system, the software used to analyze the communication, coding procedure and the reliability of the coding are introduced.

3.5.3.1 Video Selection

The communication analysis focused only on the image related videos, as the videos were very long (around one and a half hours each) and it would have been time consuming to analyze the 64 whole videos. The benefit of analyzing parts of the video, not the whole video, was that it was easy to get a connection between the coded behaviors and the tasks. For the following reasons, only the parts of the videos related to images were analyzed in this study:

1) Images are in the Clip Art folder in Microsoft Word. The usability of the wedding invitation Clip Art is the main focus in this study.
2) Images are important for a wedding in all cultures (Clemmensen, Shi, Sun, & Yammiyavar, 2008). The pilot study (Clemmensen, Shi et al., 2008) shows the background is not necessary for Danish users.
3) Images are understandable for both local and foreign evaluators. Since the texts are in the local language and foreign evaluators cannot understand it, there may be bias in analyzing the parts of the video related to the text. People with different cultural backgrounds may have different interpretations of the images (Horton, 2005), and thus it shows the communication differences between the local and distant pairs.
4) The tasks of selecting and editing the images are close to the end of the test. Users have experience in doing thinking aloud.
5) More effort is spent on preparing the materials of images than on preparing the materials of backgrounds and texts both in the pilot study and in the experiment. In this study, the images are selected by doing card sorting (Nawaz & Clemmensen, 2007).

Windows Movie Maker was used to cut down the videos and keep the image related videos. The average time of the selected videos was around 12 minutes.

3.5.3.2 Coding System

Codes are abstractions of the data (Fisher & Sanderson, 1996). They can be seen as labels attached to data elements in order to get “the meaning of the data while reducing the variability of its vocabulary” (Fisher & Sanderson, 1996, p. 29). In order to investigate the communication between the evaluator and user, their behaviors should be coded based on the themes to be examined. Since detecting usability problems was the main purpose of the usability tests, in order to see the evaluators' and users' communication patterns for finding the problems, the problems were coded as the state event with duration of time. Then the verbal behaviors could be coded as point events to examine what kind of verbal behaviors the evaluators and users had. In order to do the coding, a coding system was developed first.

The coding system was used to measure the behaviours that were of interest for this research and could be observed in the research setting (Cozby, 2003). Such a system should be simple and clear in order to help the coders easily categorize the behaviors (Cozby, 2003, p. 105). The coding system in this study was developed based on the discussion in section 2.3.1.2, and is described below (see the detailed explanation of the codes in Appendix 14):

State Events

1. **State events of usability problems**: Communication in the usability problem finding period is very important; accordingly, the problems are coded as state events.

    1) UPU: the problem is started by the user's current behaviour or speech.
    2) UPE: the problem is started by the evaluator's directed probing behaviours, without any negative comments from the user or without any problems of the user's task performing.

Two modifiers are with UPU and UPE: Modifiers are the codes that can be attached to behaviours and subjects in order to "limit the scope of the behaviour or to specify the subject more precisely" (Observer, 2005, p. 96).

A. Usability problem number: each problem related to the tasks of images is given a number. (The usability problems were those written down by the evaluators after each test.) In each test, the problems are labelled from 1 to x (x is the number of the problems, if there are 9 problems related to the images, then x is 9). By giving each problem a number, we could easily see what happened within the problem finding period.

    a. UP 1
    b. UP 2
    c. UP 3

d. UP 4
e. UP 5
f. UP 6
g. UP 7
h. UP 8
i. UP 9
j. UP 10
k. Two UPs are discussed/ talked at the same time and hard to distinguish

Note: In the coding system, UP represents usability problem.

B. Usability problem severity: the severity is the one given by the evaluators.

a. M (minor problem)
b. I (important problem)
c. C (critical problem)

2. User's state behaviours---with 3 modifiers:

1) Classic thinking aloud (Ericsson and Simon's level 1 and 2 data, explained in Appendix 14)
2) Talk rather than classic thinking aloud data, such as comments, explanations, opinions by answering questions and so on
3) Silence (the duration of silence should be more than 10 seconds)

Point Events

1. Users' point event behaviours:

1) Asking questions related to the task
2) Short reply, such as ok, yea, no
3) Negative comments
4) Positive comments
5) Suggestions, which imply how to improve the application
6) Culture related comments: if the user mentioned in Denmark/ China…, or religions or any cultural issues, then code it as culture related comments

7) Others: any behaviour that may play a role in finding usability problems or potential problems, but not belonging to the above categories. It also includes the neutral comments, such as "it is ok."

Notice:

For comments and suggestions, even though they are broken by the evaluator's "affirmative response" or "digging dipper probing," they were coded only once if they were mentioned at the same time and no other topic was in between. If the user's talk was broken by the evaluator's affirmative sentence, then just the first sentence was coded as "negative," "positive" or "suggestion."

2. Evaluators' point event behaviours:

1) Affirmative response, the short sentence/ words to act as an active listener, such as ok, yes, en, sure...
2) Probing behaviour----modifiers (A, B, C, D):
   A. Directed probing (evaluator controlled probing): basically, all probes that are not followed with the user's current talk, behaviour or emotion.
   B. Act as reminders to think aloud or talk, such as "and now…?," "what are you thinking," "what are you looking for."
   C. Digging deeper probing (user controlled probing), including asking for clarifying or speaking out the user's meaning by guessing: basically, all probes followed with the user's current talk, behaviour or emotion.
   D. Others: including all the general questions and planned probes (Dumas & Loring, 2008) that are provided in the testing protocol.
3) Classic reminders, such as please keep talking
4) Help behaviour: One help is just coded once, even though there are many communications. The modifiers are:
   A. Try to help the user to figure out the problem--- do the help or willing to do the help
   B. No clear answer or help: do not give the user clear help or answer, but encourage the user to figure it out.
5) Others: some important things you want to show other people (take notes).

The above coding system is explained in detail in Appendix 14. Table 4 shows the coding system in a clearer way without the explanations.

Table 4: The coding system

Subject	**Behaviors**	**Modifier (groups)**	**Event type**
User	UPU	UP number (11 modifiers)	State event
		UP severity (3 modifiers)	State event
	User's state behavior	Classic thinking aloud	State event
		Talk, rather than classic thinking aloud	State event
		Silence	State event
	Asking questions		Point event
	Short reply		Point event
	Negative comments		Point event
	Positive comments		Point event
	Suggestions		Point event
	Culture related comments		Point event
	Others		Point event
Evaluator	UPE	UP number (11 modifiers)	State event
		UP severity (3 modifiers)	State event
	Affirmative response		Point event
	Probing behavior	Directed probing	Point event
		Act as reminders to think aloud or talk	Point event
		Digging dipper probing	Point event
		Others	Point event
	Classic reminders		Point event
	Help behavior	Try to help	Point event
		No clear answer or help, but encourage	Point event
	Others		Point event

Note. UP represents usability problem.

In order to better understand the coding system, Table 5 shows an example of coding UPU and Table 6 shows an example of coding UPE.

Table 5: An example of UPU

Time	Person	Speech	**Coding: UPU 1 start and end** UPU1 description and severity: pictures provided are too pink and tacky (UPU: 1m)
2:39	U	If I have to choose one or more pictures, I would probably choose one of those	
2:42	E	Ok	Affirmative

2:44	U	It is…. I guess, a little bit pink and tacky, but I would choose this one	UPU 1 (m) started Negative comment
2:5	E	Yea, sure, try	Affirmative response; UPU 1 (m) ended

Note. U for user, E for evaluator

In Table 5, the UPU is coded as 1 and m, meaning that this problem is the first usability problem that the evaluator wrote down on the form of image related tasks and it is a minor problem. From Table 5 we can see that this problem is from the users' negative comments, and it is coded as UPU. The UPU started before the user said, "It is…. I guess, a little bit pink and tacky, but I would choose this one" and afterwards the evaluator said "ok." It ended after the evaluator's affirmative response. The problem is described as "pictures provided are too pink and tacky." The user's negative comment clearly shows this problem. The users' and evaluators' verbal behaviours within the UPU period are "negative comment" and "affirmative response," respectively.

Table 6: An example of UPE

Time	Person	Speech	**Coding: UPE 8 start and end** UPE8 description and severity: pictures of other grooms and brides are not an option (UPE : 8 i)
38:35	U	I would probably avoid choosing the stuff like big hearts, and rings and those kind of stuff	Negative comments
38:44	E	What's wrong with hearts or rings?	Digging deeper
38:48	U	Well, it might be ok. But to me, it is a little bit too tacky. En…	(continue the negative comments)
39:07	E	What about the pictures with bride and groom? What do you think of that kind of pictures?	UPE 8 (i) started, Directed probing
39:16	U	En…I am just trying to get a feeling of…, well, maybe I would be interested in inserting a picture of me and my future wife	Suggestion

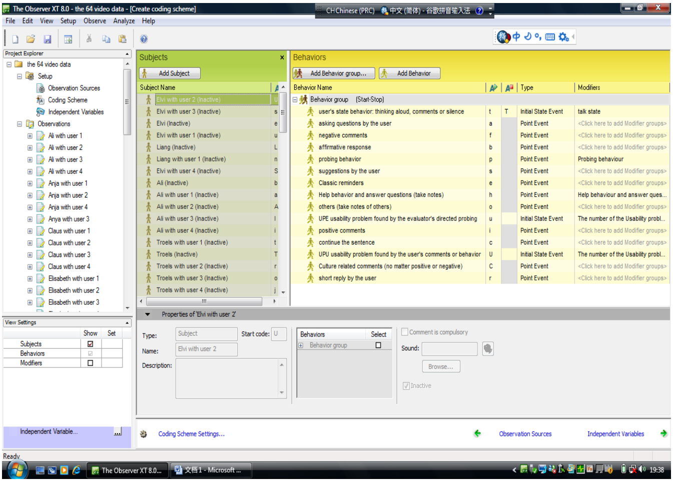
39:29	E	Ok	Affirmative
39:30	U	that would be obviously in wedding dress, but maybe just a picture of us	(continue the suggestion)
39:51	E	Yea	Affirmative (UPE 8 i ended)

Note. U for user, E for evaluator

The UPE in Table 6 is the eighth problem that the evaluator wrote down on the form and it is an important problem. This problem is started by the evaluator's directed probing, not by the user's initiated verbal or non-verbal behaviours, so it is coded as UPE. From Table 6 we can see that before the evaluator's directed probing at 39:07, they were discussing another topic. If the evaluator did not ask this question, the user may not have mentioned the problem of "pictures of other grooms and brides." The question that the evaluator asked, "What about the pictures with bride and groom?" is not a general question, such as, "What do you think of the pictures that we provided?" but a specific question, directing the users' attention to a specific part of the application. If the evaluator did not think it would be a problem, he may not have asked this specific question. Thus the problem information could be regarded as being transmitted from the evaluator to the user, though it also had to be confirmed by the user, or else it was not a usability problem.

3.5.3.3 The Software Used to Analyze the Communication

Observer XT 8.0 was used to analyze the communication. Observer has powerful functions in coding behavioral events, recording times and associating events, and sorting, organizing and analyzing the data. It has been used by many researchers to analyze the social interaction and communication patterns (Eide, Quera, & Finset, 2003; Graugaard, Holgersen, Eide, & Finset, 2005; Koch & Zumbach, 2002; Lavelli & Poli, 1998). Yammiyavar et al. (2008) used Observer to analyze Danish and Indian evaluators' and users' non-verbal communication behaviors in usability testing (section 1.6). In this research, Observer was used to analyze the evaluators' and users' verbal communication behaviors. Figure 16 and Figure 17 show the interfaces of Observer XT 8.0.



The Observer XT 8.0 - the 64 video data - [Create coding scheme]

Project Explorer: the 64 video data; Setup: Observation Sources, Coding Scheme, Independent Variables; Observations: Ali with user 1, Ali with user 2, Ali with user 3, Ali with user 4, Anja with user 1, Anja with user 2, Anja with user 4, Anya with user 3, Claus with user 1, Claus with user 2, Claus with user 3, Claus with user 4, Elisabeth with user 1, Elisabeth with user 2, Elisabeth with user 3

Subject Name	
Elvi with user 2 (Inactive)	U
Elvi with user 3 (Inactive)	s
Elvi (Inactive)	e
Elvi with user 1 (Inactive)	u
Liang (Inactive)	L
Liang with user 1 (Inactive)	n
Elvi with user 4 (Inactive)	S
Ali (Inactive)	b
Ali with user 1 (Inactive)	a
Ali with user 2 (Inactive)	A
Ali with user 3 (Inactive)	l
Ali with user 4 (Inactive)	i
Troels with user 1 (Inactive)	t
Troels (Inactive)	T
Troels with user 2 (Inactive)	r
Troels with user 3 (Inactive)	o
Troels with user 4 (Inactive)	j

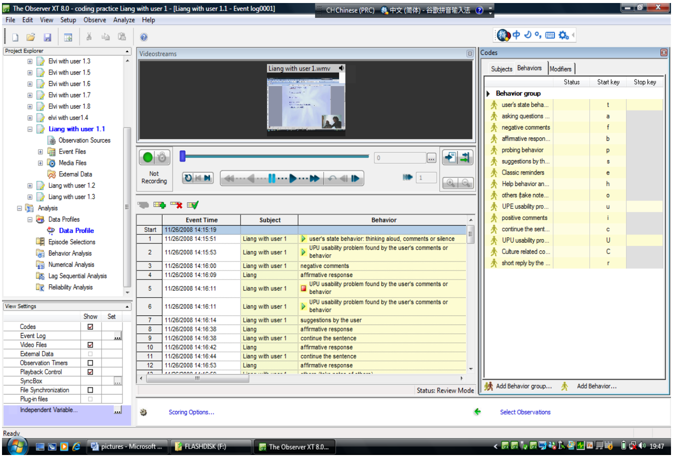
Behavior Name			Type	Modifiers
Behavior group (Start-Stop)				
user's state behavior: thinking aloud, comments or silence	t	T	Initial State Event	talk state
asking questions by the user	a		Point Event	<Click here to add Modifier groups>
negative comments	f		Point Event	<Click here to add Modifier groups>
affirmative response	b		Point Event	<Click here to add Modifier groups>
probing behavior	p		Point Event	Probing behaviour
suggestions by the user	s		Point Event	<Click here to add Modifier groups>
Classic reminders	e		Point Event	<Click here to add Modifier groups>
Help behavior and answer questions (take notes)	h		Point Event	Help behaviour and answer ques...
others (take notes of others)	o		Point Event	<Click here to add Modifier groups>
UPE usability problem found by the evaluator's directed probing	u		Initial State Event	The number of the Usability probl...
positive comments	i		Point Event	<Click here to add Modifier groups>
continue the sentence	c		Point Event	<Click here to add Modifier groups>
UPU usability problem found by the user's comments or behavior	U		Initial State Event	The number of the Usability probl...
Culture related comments (no matter positive or negative)	C		Point Event	<Click here to add Modifier groups>
short reply by the user	r		Point Event	<Click here to add Modifier groups>

Properties of 'Elvi with user 2'

Type: Subject Start code: U

Name: Elvi with user 2

Figure 16: Observer XT 8.0 used for coding the videos: define the codes



The Observer XT 8.0 - coding practice Liang with user 1 - [Liang with user 1.1 - Event log0001]

	Event Time	Subject	Behavior
Start	11/26/2008 14:15:19		
1	11/26/2008 14:15:51	Liang with user 1	user's state behavior: thinking aloud, comments or silence
2	11/26/2008 14:15:53	Liang with user 1	UPU usability problem found by the user's comments or behavior
3	11/26/2008 14:16:00	Liang with user 1	negative comments
4	11/26/2008 14:16:09	Liang	affirmative response
5	11/26/2008 14:16:11	Liang with user 1	UPU usability problem found by the user's comments or behavior
6	11/26/2008 14:16:11	Liang with user 1	UPU usability problem found by the user's comments or behavior
7	11/26/2008 14:16:14	Liang with user 1	suggestions by the user
8	11/26/2008 14:16:38	Liang	affirmative response
9	11/26/2008 14:16:38	Liang with user 1	continue the sentence
10	11/26/2008 14:16:42	Liang	affirmative response
11	11/26/2008 14:16:44	Liang with user 1	continue the sentence
12	11/26/2008 14:16:53	Liang	affirmative response

	Status	Start key	Stop key
Behavior group			
user's state beha...		t	
asking questions ...		a	
negative comments		f	
affirmative respon...		b	
probing behavior		p	
suggestions by th...		s	
Classic reminders		e	
Help behavior an...		h	
others (take note...		o	
UPE usability pro...		u	
positive comments		i	
continue the sent...		c	
UPU usability pro...		U	
Culture related co...		C	
short reply by the ...		r	

Figure 17: Observer XT 8.0 used for coding the videos: doing the coding

With Observer XT 8.0, we could define the evaluators and users and the behaviors we wanted to analyze based on the coding system (Figure 16). After defining the codes, the coding will start (figure 17). When doing the coding, we could see the video of the whole test, such as the screen, the keyboard activity of the user, the faces of the user and the evaluator and the sound (section 3.3.3). We just needed to select the person and the relevant behaviors which had been defined beforehand while doing the coding. Later, we could use this software to analyze the duration and the frequency of the codes, the behaviors within the state events and the order of the events.

3.5.3.4 The Coding Procedure

The coding procedure is:

1) Transcribe all the image related videos into Excel in order to make the coding easier and more correct, since there are some state events which require coding the starts and endings. With the transcription, we can mark the starts and endings on Excel and then code them. (Appendix 13 shows the way to transcribe the videos into Excel).
2) There are two people to do all the coding (the researcher and a Master's student). Before coding the videos, the coders needed to be trained in the same way to code the videos. Training is a very important procedure for the coding, since coding is not a natural activity (Krippendorff, 2004). The trainings included:
    a) Introduce the test and show the other coder a whole video (not only the video related to the image but a whole test) to make him familiar with the test.
    b) Introduce each item in the instruction and discuss with the other coder to make sure he understands. See the coding instruction in Appendix 14.
    c) Ask the coder to read the instruction again until he is familiar with all the items.
    d) Both coders code the same video and check the reliability.
    e) Discuss the coding together by showing the video and our coding result (confusion matrix and comparison list from Observer reliability analysis).
    f) Code the same video again together and discuss each code.
    g) Code the same video separately again after the discussion and examine the reliability again.
    h) Code the first video by the two coders together again and examine the reliability.
    i) Code the second video and check the reliability, and discuss the result.
    j) Discuss the second coding together and code the same video together.
    k) Each person codes the second video separately again and examine the reliability

l) Code the third video and examine the reliability and discuss again.
m) Code the fourth video.
n) If the reliability is good, then start to code all other videos.

3) Each person codes two tests from each evaluator, so each person has to code 32 videos. But since four videos have been coded together as the training session, each person has to code 30 more videos.
4) Before coding each video, the coders need to mark the state events' starts and ends in the transcription in order to do the coding easier and accurately in Observer.
5) When doing the coding, the two coders continually discuss the problems and uncertain issues. Most uncertain issues can be solved while coding the videos. Notes on the important issues are taken to be discussed later.
6) After finishing all the coding, the two coders discuss their codes again based on the notes taken during the coding.
7) Modify each person's own coding after the discussion.

3.5.3.5 Reliability of the Coding

The average time for the selected videos was around 12 minutes. The average coding time for each video was around 3 hours, not including the transcribing time. So the ratio between the coding time and observation time was 15:1, which is much higher than the research-oriented video coding 10:1 (Yammiyavar et al., 2008). This indicates that the two coders spent sufficient time on coding, implying that the coding used for further analysis has good quality.

In order to see whether the two coders had the same understanding of the codes and code the videos in a similar way, inter-rater reliability was used to estimate the correlation of the coding agreements (Yammiyavar et al., 2008). After sufficient trainings (described above), two coders code the same four videos. The four videos chosen to code together were from the four groups, Danish pairs (DEDU), Chinese evaluator-Danish user pairs (CEDU), Chinese pairs (CECU), and Danish evaluator-Chinese user pairs (DECU). In each group, one video was chosen to code. The inter-rater reliability was then calculated by using the Observer XT 8.0, see Table 7.

Table 7: Inter-rater reliability

Summed statistics	Video 1 DEDU Test	Video 2 CEDU test	Video 3 CECU test	Video 4 DECU test
	Value	Value	Value	Value
Agreements	96	77	70	110
Disagreements	16	24	14	11
Proportion of agreements	0.86	0.76	0.83	0.91
Kappa	0.85	0.74	0.82	0.9
Rho	0.95	0.98	0.95	1

Note. Kappa and Rho values were all significant on the 0.01 level (99% level of confidence).

Table 7 shows that all proportions of agreements are from 0.76 to 0.91 which is quite high. Table 7 also gives the Cohen kappa and Pearson values which are the most commonly used measures for consistency of data sets (Jansen, Wiertz, Meyer, & Noldus, 2003). Kappa is "the proportion of agreement after chance agreement is removed from consideration, i.e., the proportion of agreement actually attained beyond chance" (Keenan et al., 1999, p. 84). It is used to calculate the agreement between two coders (Barrett, 2001; Kundel & Polansky, 2003). The up limit of Kappa is 1 and occurs when two videos are coded exactly the same. Kappa is 0 when "the proportion of agreement equals what would be predicted by chance" (Keenan et al., 1999, p. 84). In this study, the kappa values varied from 0.74 to 0.9. According to Landis and Koch (1977), the agreements between the two coders are substantial (0.6-0.79) to almost perfect (0.8-1.00). Pearson's Rho value was used for calculating the degree of consistency between the two coders (Stemler, 2004). The results indicate that the two coders' coding was highly consistent (from 0.95 to 1). The reliability analysis show that there is high agreement between the two coders about when and what types of verbal behaviors they observed.

Before coding all the videos by the coders separately, we had a two-week training session in order to ensure that the two coders would have the same understanding of the codes. For each video, we also had a transcription version in Excel. The transcription helped to code the state events in a more accurate way. We marked all the state events and important point events in Excel before coding the video. If we had any questions about how to code some specific content, we discussed it together and reached an agreement. Even though we coded the videos separately, if we had any questions, we also discussed these questions to make sure we gave the right codes.

3.5.3.6 Summary of the Communication Analysis

The communication patterns were analyzed in two different ways: analyzing the communication patterns in the selected videos, and analyzing the communication patterns within the usability problem period. The problems are those related to the tasks of images which are evident from the selected videos. Usability problems were analyzed as UPU and UPE separately, since they are assumed to have different communication patterns. SPSS 16.0 was used to analyze the coded behaviours to see whether there were any significant differences between the Danish and Chinese evaluators' and users' communication patterns.

3.6 Chapter Summary

Chapter 3 has presented the research design, data collection procedures and data analysis in this research. An experimental study was conducted in Denmark and China to investigate the cultural influence on usability testing. The benefit of carrying out the experimental study was that we could evaluate the role that culture plays in usability testing by controlling other potential influences in different groups. Through statistical analyses, we could generalize the findings (identified usability problems and communication patterns) to the similar usability testing settings. The disadvantage of performing the experimental study was that the usability testing may not have been exactly the same as the real situation, and thus the external validity may have been affected to some extent. In order to examine the sub-research questions, usability tests were conducted in both Denmark and China with both Danish and Chinese evaluators. The way to conduct these tests and the way to analyze the data were also discussed in this chapter.

The next chapter presents the results.

4 Results

This chapter presents the findings of this research. The results include the analysis of the demographics of the evaluators and users, the comparison of the usability problems found and problem severity rated by evaluators in different cultural settings, the comparison of the problems found and problem severity rated by evaluators and users, the communication patterns of the evaluators and users and the interviews.

4.1 Demographics

The demographic information of the evaluators and users which was obtained from the background questionnaire is shown in Table 8 and Table 9.

Table 8: The evaluators' demographic information

Items	Danish evaluators (DEs) when with Danish users (DUs)	Chinese evaluators (CEs) when with Danish users (DUs)		Chinese evaluators (CEs) when with Chinese users (CUs)	Danish evaluators (DEs) when with Chinese users (CUs)
Average age	33.00 (SD=6.06)	30.25 (SD=6.65)		27.25 (SD=2.06)	32.75 (SD=1.71)
Average years of education	19.75 (SD=3.77)	19.50 (SD=1.73)		19.25 (SD=1.26)	18.25 (SD=1.26)
Nationality	Denmark 4	China 4		China 4	Denmark 4
First language	Danish 4	Chinese 4		Chinese 4	Danish 4
Grew up in	Denmark 4	China 4		China 4	Denmark 4
Gender	2 male and 2 female	1 male and 3 female		2 male and 2 female	3 male and 1 female
Years of being a usability specialist	2 DEs with more than 4 years, and 2 DEs with around 1 year	2 CEs with more than 4 years, 1 CE with 1 to 2 years, 1 CE with 2 to 4 years		3 CEs with 1 to 2 years, 1 CE with 2 to 4 years	3 DEs with 2 to 4 years, 1 DE with more than 4 years
Number of usability tests they have done	3 DEs have done more than 5 tests, 1DE has done 3 to 5 tests	3 CEs have done more than 5 tests, 1CE has done 3 to 5 tests		All 4 CEs have done more than 5 tests	All 4 DEs have done more than 5 tests
Spoken English	All four DEs are good to very good	All four CEs are good to very good		All four CEs are good to very good	All four DEs are good to very good
Official language in the office	Danish 4	Chinese 2 English 2		Chinese 2 English 2	Danish 4
Preferred language while working with computer	Danish 1 English 3	Chinese 2 English 2		Chinese 4	Danish 2 English 2
The functions with editing text	All have used all listed functions	All have used all listed functions		All have used all listed functions	All have used all listed functions
The function of	All have used it	All have used it		All have used it	All have used it

Inserting picture					
Whether have seen Danish/Chinese Wedding invitations	All DEs have seen Danish wedding invitations	All CEs have never seen Danish wedding invitations		All CEs have seen Chinese wedding invitations	All DEs have never seen Chinese wedding invitations

From Table 8, we can see that when doing tests with Danish users, two male and two female Danish evaluators, as well as one male and three female evaluators, participated in this research. The average age of Danish evaluators was 33, and the average age of Chinese evaluators was 30. The average number of years of education was similar (19.75 vs. 19.50). Mann-Whitney U test which is a nonparametric equivalent of the independent samples t test (Hinton et al., 2004) was used to examine whether the age and the years of education have significant differences between the Danish and Chinese evaluators when with Danish users. Mann-Whitney U test can be used for samples of N<20 in each group. In this study, the samples of both Danish and Chinese evaluators have "N=4." The results show that there is no significant difference between Danish and Chinese evaluators' age and years of education (U_{age}=6.000, p_{age}=0.554; $U_{edu.}$=7.000, $p_{edu.}$=0.770). The number of years of experience of being a usability specialist was also similar. Two Danish and two Chinese evaluators had more than four years of experience working as usability specialists, the other two Danish evaluators had around one year experience, and the other two Chinese evaluators had around one to four years. However, the usability tests show that Danish and Chinese evaluators had similar experience in doing usability tests: three Danish evaluators and three Chinese evaluators had done more than five tests before, whereas one Danish evaluator and one Chinese evaluator had done three to five tests before.

In the testing with Chinese users, two male and two female Chinese evaluators, as well as three male evaluators and one female Danish evaluator, participated in the study. The average age of the Chinese evaluators was 27 and the average age of the Danish evaluators was 33. The Chinese evaluators were younger than the Danish evaluators were (U< 0.001, p=0.020). Age differences may, however, not play a big role in facilitating the tests and finding usability problems, since all evaluators were adults, neither too old, nor too young. There was no significant difference between the Danish and Chinese evaluators' years of education (U=4.000, p=0.234). Regarding the number of years of experience of being a usability specialist, Table 8 shows that both Danish evaluators and Chinese evaluators had more than one year experience of being usability specialists and all had done more than five usability tests before. Thus, they could be regarded as experienced usability specialists.

All evaluators could speak English well. Some evaluators used English as the official language in their offices, and others used English as their preferred language while working with computers. English was not far removed from the Danish and Chinese evaluators' life. Further, they had similar and good experience in using MS Word. All local evaluators had seen wedding invitations in their own cultures, and all foreign evaluators had never seen the wedding invitations of the target cultures.

The evaluators' demographic information tells us the local and foreign evaluators in this study are comparable. If they find different usability problems, this may be because of cultural issues, not other issues. The users' demographic information is shown in Table 9.

Table 9: The users' demographic information

Items	Danish users (DUs) when with DEs	Danish users (DUs) when with CEs		Chinese users (CUs) when with CEs	Chinese users (CUs) when with DEs
Average age	27.25 (SD=4.47)	27.00 (SD=5.53)		24.31 (SD=1.07)	23.69 (SD=1.30)
Average years of education	17.75 (SD=2.84)	16.67 (SD=2.09)		17.06 (SD=1.48)	17.19 (SD=1.51)
Nationality	Denmark	Denmark		China	China
First language	Danish	Danish		Chinese	Chinese
Grew up in	Denmark	Denmark		China	China
Gender	11 male, 5 female	7 male, 9 female		6 male, 10 female	7 male, 9 female
Spoken English	All good to very good	All good to very good		5 normal, 11 good to very good	3 normal, 13 good to very good
Official language in the office	Danish 8 English 8	Danish 10 English 6		Chinese 13 English 3	Chinese 12 English 4
Preferred language while working with computer	Danish 7 English 9	Danish 7 English 9		Chinese 11 English 5	Chinese 12 English 4
The functions with editing text	15 DUs used all functions, and 1 used most of the functions	12 DUs used all functions, and 4 used most of the functions		7 CUs used all functions, and 8 used most of the functions	8 CUs used all functions, and 7 used most of the functions
The function of Inserting picture	14 users have used it	All have used it		All have used it	15 users have used it
Whether have seen Danish/Chinese Wedding invitations	15 DUs have seen the Danish invitation	15 DUs have seen the Danish invitation		13 CUs have seen the Chinese invitation	13 CUs have seen the Chinese invitation

The Mann-Whitney U test shows that Danish users with local and foreign evaluators had similar ages (U=109.500, p=0.484), and that Chinese users also had similar ages with local and foreign evaluators (U=85.00, p=0.091). The average years of education were similar for users in different groups (U_{DU}=96.500, p_{DU}=0.345; U_{CU}=125.000, p_{CU}=0.906). All Danish users thought they could speak English well, and most Chinese users also thought they could speak English well; however, 8 Chinese users thought their spoken English was normal. But since Chinese users were selected through an English interview and all had also passed some standardized English tests (such as TOEFL or IELTS, see section 3.2.2), they were able to express their thoughts in English, albeit their English skills may not have been as good as that of the Danish users. Since the Chinese users with local and foreign evaluators had similar English skills, the results of comparing the two groups of Chinese users should not have been influenced. The Danish users with different evaluators had similar experience of using the functions of editing text and inserting pictures. Even though there were fewer Chinese users than Danish users who had used all the functions of editing text, there was no difference between the Chinese users with local and foreign evaluators. Most users had seen invitations in their own cultures. Thus, users with local and foreign evaluators were also comparable in this study.

The demographic information shows that the evaluators and users in Denmark and China were qualified and comparable in different cultural settings; accordingly, the internal validity of the study is high and there are almost no task-specific confounding variables.

As discussed in section 3.1.2.1 and section 3.2.1, this research is mainly based on existing culture theories and their findings which are introduced in chapter 2. The participants were local people and they were supposed to have the tendency described by researchers (i.e., Nisbett and his colleagues and Hall). In the background questionnaire (Appendix 2 and 3), only two short questions were used as detectors to check the participants' cultural tendency. The results are given in Table 10.

Table 10: Categorization orientations

Item	Categorize cow, chicken and grass		Target object belongs to group 1 or 2	
Reason of Categorization	Based on relation (cow and grass)	Based on features (cow and chicken)	Based on family resemblance (G1)	Based on Rules (G2)
Danish participants	35%	65%	45%	55%
Chinese participants	60%	40%	100%	0

Note. 20 Danish and 20 Chinese participants answered the two questions.

From Table 10, we can see that Chinese participants had the tendency of grouping objects according to the relation and the family resemblance as described by Nisbett and his colleagues (section 2.2.2.2), whereas Danish participants had the opposite tendency. They tended to group the objects according to the features of the objects (such as both animals) and the rules (such as with straight stem).

The above results show that Danish and Chinese participants in this study are comparable and, to some extent, represent the target people described by culture theories. The usability problem analysis and communication analysis follow next.

4.2 Comparison of the Usability Problem Found and Problem Severity Rated by Evaluators in Different Cultural Settings

This section presents the usability problem found and problem severity rated by evaluators in different cultural settings, involving the investigation of hypothesis 1 and hypothesis 2 put forward in section 2.3.2.1.

4.2.1 Analysis of the Usability Problems Found by Evaluators in Different Cultural Settings (Hypothesis 1)

Hypothesis 1: The usability problems found by local evaluators are different from the usability problems found by foreign evaluators.

- H1 a (in Denmark): When with Danish users, the usability problems found by Danish evaluators are different from the usability problems found by Chinese evaluators.
- H1 b (in China): When with Chinese users, the usability problems found by Chinese evaluators are different from the usability problems found by Danish evaluators.

As introduced in section 2.3.2.1, hypothesis 1 is divided into two sub-hypotheses, H1a and H1b. Since the two sub-hypotheses are analyzed in the same way, they will be analyzed together, but confirmed or rejected individually. If they are both supported by the data, or if they are both rejected, then hypothesis 1 is confirmed or rejected. If one sub-hypothesis is supported and the other is rejected, we cannot give a simple answer that hypothesis 1 is confirmed or rejected. The answer should be based on the cultural settings, i.e. with Danish or Chinese users.

In this study, evaluators were asked to write down the usability problems after each test. Table 11 shows the number of problems found by each evaluator after each test.

Table 11: The number of usability problems found by evaluators

Groups	Evaluators	Test 1	Test 2	Test 3	Test 4	**Total**
Danish evaluators with Danish users (DEDU)	Danish evaluator 1	18	13	13	20	64
	Danish evaluator 2	8	13	18	10	49
	Danish evaluator 3	13	15	25	14	67
	Danish evaluator 4	6	7	3	4	20
	Total	45	48	59	48	**200**
Chinese evaluators with Danish users (CEDU)	Chinese evaluator 1	15	5	8	12	40
	Chinese evaluator 2	17	17	19	9	62
	Chinese evaluator 3	8	15	14	4	41
	Chinese evaluator 4	16	13	16	12	57
	Total	57	50	57	36	**200**
Chinese evaluators with Chinese users (CECU)	Chinese evaluator 1	17	13	7	8	45
	Chinese evaluator 2	15	11	7	14	47
	Chinese evaluator 3	14	7	13	6	40
	Chinese evaluator 4	18	9	14	7	48
	Total	64	40	41	35	**180**
Danish evaluators with Chinese users (DECU)	Danish evaluator 1	20	13	12	9	54
	Danish evaluator 2	11	10	2	6	29
	Danish evaluator 3	16	26	10	12	64
	Danish evaluator 4	17	11	15	10	53
	Total	64	60	39	37	**200**

When with Danish users, Danish and Chinese evaluators found the same number of usability problems (200). When with Chinese users, Danish evaluators found even more problems than did the Chinese evaluators. But there was no significant difference (t=0.730, sig=0.471). Hence, there was no significant difference in the number of usability problems found by local and foreign evaluators, regardless of whether they were with Danish or Chinese users.

The problems shown in Table 11 are actually "problem instances" (section 3.5.1.1). As discussed in section 3.5.1, the problems written down by evaluators were categorized into non-overlapping usability problem types. A new problem list was developed on the basis of the original problem list and by considering the categorization rules introduced in section 3.5.1.1 (see section 3.5.1.2). Table 12 shows the new problem list.

Table 12: The new problem list

UPID	**Usability Problems**
UP1	Some images are not used for Danish/Chinese wedding invitations, but may be used for wedding invitations in other cultures.
UP2	Inappropriate images: such as too old fashioned, or images are not suitable for wedding (such as images should not convey religious matter), or need more traditional images…
UP3	Image collection aesthetic problem: such as images are not pretty/ romantic/ good for wedding invitations, or need more images with flowers or rings.
UP4	Image resolution/sizes problem: such as too low resolution, when it is enlarged, it is not clear; or do not know the resolution of the image or image thumbnail; or the sizes of the images are not the same.
UP5	Image collection range problem (too few): such as the images in the image folder are not enough to select from.
UP6	Image moving problem (in Word), which is the problem related to move images on Word: such as hard to move images or cannot put it on the suitable place.
UP7	Problems with the three folders that we provided, which are problems related to find/ open/ deal with /understand the folders of images, texts or backgrounds.
UP8	Image preview problems in the clipart folder: such as image thumbnail size is small; do not know how to preview the images; cannot open it by double clicking it; cannot enlarge preview size……
UP9	Problems with inserting images/ texts from the clipart to Word: such as not convenient to do that, not know how to do that.
UP10	The image, text and background layout problem (in word): such as problems with text wrapping tool; hard to make the text over the image; do not know how to make a picture as a background ….
UP11	Image formatting problem: such as editing size, contrast, brightness, colour and so on.
UP12	Clipart match problem: such as the background and images are not matched; do not know how to make the background of the images transparent; would like to remove the background from the image; cannot change the colour of the image background.
UP13	Concept problem of the application problem: such as not necessary to use this application, or would like to have a more powerful application, or Word is not suitable for making wedding invitation.
UP14	Problems of no category of images/texts/ backgrounds: the images/ texts/backgrounds should be categorized into some sub-categories, or design and classify the images/ texts/ backgrounds/ for different people or according to the themes of the backgrounds.
UP15	Inappropriate invitation texts, which means the content/expression of the invitation is not appropriate for Danish/Chinese wedding invitations (not for the culture, or not for the wedding, or old fashioned expression…)

UP16	Invitation text quality problems: the text may be required more consideration to make it better.
UP17	It is not easy to choose the appropriate font(s) for the text: such as the text font, the font style, the text color and the size of the words.
UP18	Hard to put the texts on the suitable place or hard to do the format for the text: such as the space between the words, the place of the text on the paper……
UP19	Inappropriate invitation backgrounds: such as some backgrounds are not for Danish/Chinese wedding invitations (not for the culture, or not for the wedding, or old fashioned or not suitable for wedding invitations).
UP20	Background quality problems: backgrounds are not pretty/ good enough.
UP21	Background editing problem: such as it took some time to edit the colour, brightness, contrast and so on.
UP22	Background collection range problem (too few): such as the backgrounds are not enough.
UP23	Background preview problem or there is no name for each background document: cannot preview the backgrounds; should open files one by one; accidently close wanted background; or too many documents are open and they collapsed on the toolbar.
UP24	Problem of making the text vertical: want to make the text vertical, but it is not easy.
UP25	Problem with the impersonal images/ backgrounds/ texts: the images/ backgrounds/ texts are not personal, pictures should not be strangers, need to be unique and original, need the function to add own photos….
UP26	Page setup problem of the invitation: such as the page size and colour setup; or should support printing on other than White A4; background should be the whole page; can not make it as folder, or from portrait to landscape.
UP27	Task support problems: such as it should provide finished examples of the invitation; or it is better to use a guide to do the letter, one step after one step; or some tasks are not necessary to do.
UP28	Would like to use google to find more images/ texts/ backgrounds, or the company should provide some online package of cliparts.
UP29	Would like to see effects of changes immediately (without clicking ok).
UP30	General Word processing problem or computer system problem: such as hard to find the picture toolbar; not happy with some functions of Word; the computer system is too slow...

Note. UP is for "usability problem".

Table 12 shows the description of the usability problems. 30 problems were extracted from the tests. All problem instances were categorized into 30 usability problems. Table 13 shows the number of problem instances within each usability problem.

Table 13: The number of the problem instances written down by evaluators for each non-overlapping usability problem in different groups

UPID	DE-DU	CE-DU		CE-CU	DE-CU
UP1	1	1		0	3
UP2	3	7		3	9
UP3	10	7		2	5
UP4	6	7		1	3
UP5	2	5		4	1
UP6	6	4		6	7

UP7	6	1		1	3
UP8	4	8		11	12
UP9	8	6		4	2
UP10	8	8		1	11
UP11	2	6		4	7
UP12	9	3		5	2
UP13	4	2		0	8
UP14	1	1		10	0
UP15	6	4		17	16
UP16	16	17		12	8
UP17	3	9		17	15
UP18	6	6		8	10
UP19	16	8		13	10
UP20	24	30		11	12
UP21	10	11		10	12
UP22	1	4		10	4
UP23	10	8		7	4
UP24	0	0		0	2
UP25	24	17		2	7
UP26	5	6		1	7
UP27	1	3		4	0
UP28	1	4		2	0
UP29	1	1		5	7
UP30	6	6		9	13
Total number of UP instances	200	200		180	200
Total number of non-overlapping UPs	29	29		27	27
Shared UPs when with users from the same culture	29	29		24	24
Unique UPs when with users from the same culture	0	0		3	3

Notes: DE-DU means Danish evaluators when with Danish users; CE-DU means Chinese evaluators when with Danish users; CE-CU means Chinese evaluators when with Chinese users; DE-CU means Danish evaluators when with Chinese users; UP means usability problem.

Most non-overlapping usability problems were identified by the local and foreign evaluators with both Danish and Chinese users. The data in Table 13 is the number of the problem instances identified in the 16 tests conducted by the 4 evaluators in each cultural setting, and thus most problems were able to be detected (Law & Hvanneberg, 2002; Nielsen & Landauer, 1993). For each test, the number of non-overlapping problems may be varied. Table 14 to Table 17 show the number of non-overlapping problems found in each test.

Table 14: The number of non-overlapping usability problems found by Danish evaluators when with Danish users

DE-DU	Test 1	Test 2	Test 3	Test 4	Total
Danish evaluator 1	14	10	8	14	23
Danish evaluator 2	6	9	11	8	18
Danish evaluator 3	9	6	15	8	20
Danish evaluator 4	6	5	2	3	13
Total					29

Table 15: The number of non-overlapping usability problems found by Chinese evaluators when with Danish users

CE-DU	Test 1	Test 2	Test 3	Test 4	Total
Chinese evaluator 1	13	5	8	10	20
Chinese evaluator 2	12	10	11	6	20
Chinese evaluator 3	8	9	10	4	19
Chinese evaluator 4	13	11	12	10	23
Total					29

Table 16: The number of non-overlapping usability problems found by Chinese evaluators when with Chinese users

CE-CU	Test 1	Test 2	Test 3	Test 4	Total
Chinese evaluator 1	12	9	6	8	18
Chinese evaluator 2	11	9	6	9	17
Chinese evaluator 3	11	7	10	6	18
Chinese evaluator 4	10	8	12	6	18
Total					27

Table 17: The number of non-overlapping usability problems found by Danish evaluators when with Chinese users

DE-CU	Test 1	Test 2	Test 3	Test 4	Total

Danish evaluator 1	13	10	10	9	22
Danish evaluator 2	10	8	2	6	17
Danish evaluator 3	14	14	8	8	23
Danish evaluator 4	7	9	8	9	19
Total					27

From Tables 14 to Tables 17, we can see that even though the number of usability problems found by evaluators in each cultural setting is similar, the number of problems found by different evaluators seems different. In order to examine the local and foreign evaluators' capacity for usability problem detection, detection rate (Hvannberg et al., 2007; Law & Hvanneberg, 2004) was calculated.

1) Detection rate

The detection rate is calculated as:

Detection rate P = Average of $|P_i| / |P_{all}|$

Table 18 shows each evaluator's mean detection rate and the average mean detection rate within each cultural setting and within each country. P_{within} is the detection rate within each cultural setting (DEDU/CEDU/CECU/DECU) and $P_{overall}$ is the detection rate when with the target users (when with Danish users or Chinese users). P_{within} means the average of the sets of problems identified in each test divided by the number of problems found in each cultural setting. $P_{overall}$ means the average of the sets of problems identified in each test divided by the number of problems found by all the local and foreign evaluators when doing the tests with Danish users (or Chinese users). Each evaluator's mean detection rate was calculated by the average of the four sets of problems identified in each test, divided by the number of problems in the related group/ country.

Table 18: Detection rates within each cultural setting and within each country

Cultural settings	P_{within}: Detection rate for each evaluator within group DEDU/CEDU/CECU/DECU	$P_{overall}$: Detection rate for each evaluator within each country (Denmark/China)
DEDU 1 (4)	0.40	0.40
DEDU 2 (4)	0.29	0.29

DEDU 3 (4)	0.33	0.33
DEDU 4 (4)	0.14	0.14
DEDU (16)	0.29	0.29
CEDU 1 (4)	0.31	0.31
CEDU 2 (4)	0.34	0.34
CEDU 3 (4)	0.27	0.27
CEDU 4 (4)	0.40	0.40
CEDU (16)	0.33	0.33
CECU 1 (4)	0.32	0.29
CECU 2 (4)	0.32	0.29
CECU 3 (4)	0.31	0.28
CECU 4 (4)	0.33	0.30
CECU (16)	0.32	0.29
DECU 1 (4)	0.39	0.35
DECU 2 (4)	0.24	0.22
DECU 3 (4)	0.41	0.37
DECU 4 (4)	0.31	0.28
DECU (16)	0.34	0.30

Note. DEDU means Danish evaluator-Danish user pairs; CEDU means Chinese evaluator-Danish user pairs; CECU means Chinese evaluator-Chinese user pairs; DECU means Danish evaluator-Chinese user pairs.

Table 18 shows that when with Danish users, the Danish evaluators' mean detection rate $P_{overall}$ is 0.29, and the Chinese evaluators' mean detection rate $P_{overall}$ is 0.33. There is no significant difference (t=0.966, p=0.342). When with Chinese users, Chinese evaluators' mean detection rate $P_{overall}$ is 0.29, and Danish evaluators' mean detection rate $P_{overall}$ is 0.30. There is thus no difference (t=0.339, p=0.737). There is no difference between the local and foreign evaluators' problem finding regardless of whether they were with Danish or Chinese users.

Figure 18 shows each evaluator's detection rate, and indicates that the detection rates of the Chinese-Chinese pairs were similar. The detection rates within the other pairs (DEDU, CEDU and DECU) were different; especially the detection rates of the Danish-Danish pairs were quite different. This implies that there may be individual difference for evaluators within each group. In each group, there were four evaluators, and each evaluator did four tests. In order to examine the detection rate differences between the four evaluators, the nonparametric test, Kruskal-Wallis test was used, since there was too small a number of data, and the parametric test ANOVA could not be used (Hinton et al., 2004, p. 262). Kruskal-Wallis test shows that Danish evaluators when with Danish users find significantly different number of problems (x^2=8.849, p=0.031). The evaluators' detection rates in other groups have no significant difference (x_{CEDU}^2=4.949, p=0.176; x_{CECU}^2=0.074, p=0.995; x_{DECU}^2=5.350, p=0.148).

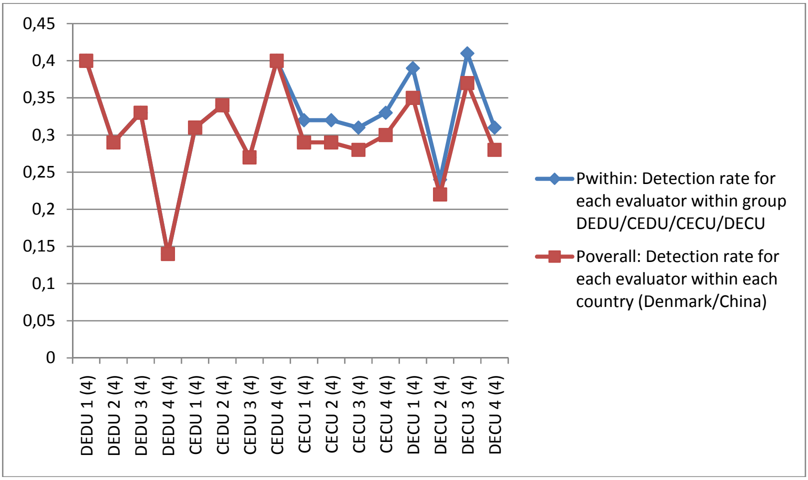


Figure 18: Each evaluator's detection rates within different groups

Note. DEDU means Danish evaluator-Danish user pairs; CEDU means Chinese evaluator-Danish user pairs; CECU means Chinese evaluator-Chinese user pairs; DECU means Danish evaluator-Chinese user pairs.

The detection rate can show only whether there is any difference in the number of problems found by evaluators, but it does not indicate the type of problems that were found by evaluators. The following sections analyze the unique problems and shared problems by evaluators in different cultural settings.

2) The unique usability problems

In the tests with Danish users, there was no unique problem which was found only by local or foreign evaluators. All problems that were found by local evaluators were also found by foreign evaluators.

Regarding the tests with Chinese users in China, local and foreign evaluators found six unique problems totally, among which local evaluators found three and foreign evaluators found the other three. Table 19 shows the unique problems found by Chinese evaluators.

Table 19: The three unique problems found only by Chinese evaluators when with Chinese users

UPID	Problem descriptions and the number of evaluators and tests with the problems
UP14	Problems of no category of images/texts/ backgrounds: the images/ texts/backgrounds should be categorized into some sub-categories, or design and classify the images/ texts/ backgrounds/ for different people or according to the themes of the backgrounds (three evaluators found this problem in eight tests)
UP27	Task support problems: some tasks are not necessary to do (two evaluators found this problem in three tests)
UP28	Would like to use google to find more images/ texts/ backgrounds, or the company should provide some online package of cliparts (two evaluators found this problem in two tests)

Note. UP is for “usability problem”.

From Table 19, we can see that most Chinese evaluators found the usability problem called “problem of no categorization of images/ texts/ backgrounds.” When checking the videos of the tests, we found that many Chinese users suggested that the images should be categorized into sub-categories, such as categorizing the images according to the themes of the backgrounds. They also suggested designing and categorizing the images for different people, since different people may like different kinds of invitations, such as young people wanting modern ones, and older people preferring the traditional ones.

In fact, the Danish evaluator 2 also mentioned the categorization issues, but she did not think they were problems. She wrote down the issues as “1. no problems, but the user suggested that the backgrounds could be improved by introducing categories of style” “2. no problems, but the texts could also be improved with a categorization of text-styles like: western, traditional and different portions/ sections of the text,” “3. no comments to image style- but images could be categorized.” The researchers interviewed the evaluator about these issues that she had written down, and asked her why she did not respond to the severity of the problems. The evaluator said she thought those were not problems, but suggestions from users showed that there would be improvement if in the future there were more backgrounds, texts or images.

From the problem communication analysis in section 4.4, we can see that many problems came from the users’ suggestions. The problem of categorization was often from the users’ suggestions, and the Chinese evaluators would take them as problems. Checking the severity of the 10 problem instances related to this problem, we found that no problem instance written down by evaluators was considered to be a minor problem, 7 instances were rated as being important, and 3 as critical. This implies that it was an important problem for the Chinese people, and thus it was easy for the Chinese evaluators to detect this problem. On the other hand, for Danish people, the categorization was not a problem. From Table 13 we observe that in the tests with Danish users, there were only two problem instances related to this problem: one was found by the Danish evaluator, and the other by the Chinese evaluator. It thus shows that categorization is not an important issue for Danish people, but may be an important issue for Chinese people, since it was harder for Danish evaluators than it was for Chinese evaluators to detect this problem.

UP 27 and UP 28 were also uncommon problems for Danish people. Table 13 shows that in the tests with Danish users, two problems were found by only one Danish evaluator in one test. The results imply that if a problem is an uncommon problem in the evaluators' culture, then it may be difficult for the evaluator to detect it when doing tests in another culture.

The unique problems found only by Danish evaluators while with Chinese users are shown in Table 20.

Table 20: The three unique problems found only by Danish evaluators when with Chinese users

UPID	Problem descriptions and the number of evaluators and tests with problems
UP1	Some images are not used for Danish/Chinese wedding invitations, but may be used for wedding invitations in other cultures (three evaluators found this problem in three tests)
UP13	Concept problem of the application problem: such as not being necessary to use this application, or would like to have a more powerful application, or Word is not suitable for making wedding invitation. (two evaluators found this problem in five tests)
UP24	Problem of making the text vertical: wanting to make the text vertical, but it was not easy. (one evaluator found this problem in two tests)

Note. UP is for "usability problem".

When with Chinese users, Danish evaluators also found three unique problems. The unique problem No. 1 and No. 24 were actually quite related to the Chinese culture. Why were they found only by foreign evaluators, and not by the local ones? UP1 was about "some images are not used for Chinese wedding invitation, but may be used for wedding invitations in other cultures." UP 24 was about "problem of making the text vertical." Nowadays, Chinese weddings have changed a lot. Many young people prefer modern wedding images which are more similar to Western wedding images. Chinese evaluators may have thought it was not a problem to provide Western wedding images since they knew many young people actually liked those.

However, when users mentioned that they preferred the traditional wedding images, not the Western images, the Danish evaluators may have thought that providing Western images was a problem.

When with local evaluators, the Chinese users might just focus on the task, which was to make a wedding invitation using the application. They might mention some good and bad aspects of the application, but it was not necessary to mention “culture.” The instruction for the evaluator was to do the tests as they usually did. In normal usability tests, people seldom mention “cultural issues” if nobody emphasizes it. However, when with the foreign evaluator, the foreign evaluator himself/ herself may have acted as a stimulator to make the Chinese user think of some cultural issues. Thus, some users may have mentioned that some images were not “Chinese,” or may have done some tasks to show the traditional Chinese wedding invitation, which may be helpful in eliciting the two unique problems.

3) The shared usability problems

Shared usability problems are those detected by more than one group of usability tests (Law & Hvanneberg, 2004). Some shared problems are shared by two evaluators, and some are shared by all evaluators. From the previous analysis of the unique problems, we can see that most problems were shared by both local and foreign evaluators. Only six unique problems were found when with Chinese users. It was therefore not necessary to analyze the problems shared by local and foreign evaluators. In this section, the shared problems within each cultural setting are analyzed. In each cultural setting, there are some unique problems and some shared problems. For example, within Danish-Danish pairs, there are six problems found only by one evaluator (Table 13). Table 21 shows the number of shared problems within each cultural setting.

Table 21: The number of shared problems within each group

	shared by all the 4 Es	shared by 3 Es	shared by 2 Es	Found Only by one E in the group
DEDU	7	7	8	6
CEDU	12	6	6	5
CECU	5	12	5	5
DECU	11	8	5	3

Note. DEDU means Danish evaluator-Danish user pairs; CEDU means Chinese evaluator-Danish user pairs; CECU means Chinese evaluator-Chinese user pairs; DECU means Danish evaluator-Chinese user pairs; E means evaluator.

From Table 21, we can see that foreign evaluators (CEDU, DECU groups) tend to have more problems shared by all four evaluators (12 vs. 7; 11 vs. 5), and fewer problems found by only one evaluator (5 vs. 6; 3 vs. 5). Figure 19 shows the number of shared problems in each group in a much clearer way.

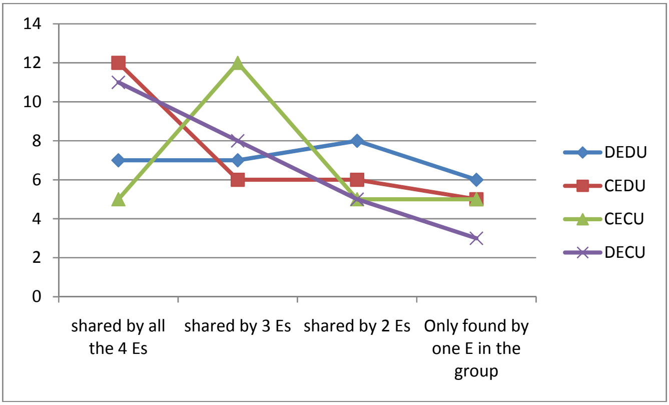


Figure 19: The number of shared problems within each group

Note. DEDU means Danish evaluator-Danish user pairs; CEDU means Chinese evaluator-Danish user pairs; CECU means Chinese evaluator-Chinese user pairs; DECU means Danish evaluator-Chinese user pairs; E means evaluator.

From Figure 19, we can see that there are more problems found by all four foreign evaluators than those found by local evaluators in both Denmark and China. The finding indicates that the problems found by foreign evaluators are more consistent than the problems found by local evaluators.

Summary of Hypothesis 1

Hypothesis 1: The usability problems found by local evaluators are different from the usability problems found by foreign evaluators.

- H1 a (in Denmark): When with Danish users, the usability problems found by Danish evaluators are different from the usability problems found by Chinese evaluators.

- H1 b (in China): When with Chinese users, the usability problems found by Chinese evaluators are different from the usability problems found by Danish evaluators.

Neither H1a nor H1b was supported by the statistic analysis, and therefore Hypothesis 1 was rejected. There was no difference between the problems found by local and foreign evaluators. However, from the analysis of the unique problems and shared problems, we know that even though the statistical analysis does not show significant differences between the problems found by local and foreign evaluators, the evaluators' cultural backgrounds may still have played a role in finding usability problems. The unique problem analysis indicates that evaluators' own thoughts may have helped in finding usability problems, and the shared problem analysis shows that the way of identifying usability problems by local and foreign evaluators seems to be different. The problem severity rating behavior is discussed next.

4.2.2 Analysis of the Danish and Chinese Evaluators' Tendency in Rating the Severity of the Usability Problems (Hypothesis 2)

Hypothesis 2: Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

- H 2 a (in Denmark): When with Danish users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.
- H 2 b (in China): When with Chinese users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

The analysis of the severity rating behavior is based on the problem instances, not the problems on the list (section 3.5.1.3). The problem instances were written down by evaluators and the severity ranks were given by evaluators. If using the problems on the list, the researcher had to decide on the severity of the problems based on the severity of all the instances given by evaluators. Since different evaluators gave different severities to the instances, it was challenging for the researcher to decide whether the problems on the list were minor, important or critical. On the other hand, if using the severity ranks given by evaluators for the problem instances directly, it was more objective to examine the evaluators' severity rating behavior. Table 22 shows the severity ranks given by evaluators.

Table 22: The severity ranks of all the problem instances given by evaluators in different groups

Evaluator- user pairs	Number of Minor UP	Number of Important UP	Number of Critical UP	Total number	Chi-Square	Sig.
DE-DU	78	62	60	200	2.920	0.232
CE-DU	73	85	42	200	14.770	0.001
CE-CU	57	95	28	180	37.633	0.000
DE-CU	71	82	47	200	9.610	0.008

Notes. DE-DU means Danish evaluators when with Danish users; CE-DU means Chinese evaluators when with Danish users; CE-CU means Chinese evaluators when with Chinese users; DE-CU means Danish evaluators when with Chinese users; UP means usability problem.

From Table 22, we can see that Danish evaluators when with Danish users rated similar numbers of minor, important and critical problems. The number of minor problems is larger than the number of important and critical problems, but Chi-Square test shows the difference is not significant (78 vs. 62 and 60; x^2=2.920, p=0.232). Regarding the other pairs, critical problems are significantly fewer than are important and minor problems. In Table 22, it is difficult to determine whether minor problems are significantly less than important problems since for the Chi-square test, if even one category is significantly different from that of other categories, it still shows significance. In order to see whether the important problems are significantly more than both minor and critical problems, Chi-square test is used for further analysis. Table 23 compares each two problem severity ranks to see whether the difference is significant between every severity rating pair.

Table 23: Comparing the difference of each two severity ranks

Evaluator- user pairs	Number of Minor UP	Number of Important UP	Number of Critical UP	Chi-Square	Sig.
Chinese Evaluators with Danish users	73	85		0.911	0.340
	73		42	8.357	0.004
		85	42	14.559	0.000
Chinese Evaluators with Chinese users	57	95		9.500	0.002
	57		28	9.894	0.002
		95	28	36.496	0.000

Danish Evaluators with Chinese users	71	82		0.791	0.374
	71		47	4.881	0.027
		82	47	9.496	0.002

Note. UP is for “usability problem”.

Table 23 shows that when Chinese evaluators are with Chinese users, the important problems are significantly more than minor and critical problems (p=0.002; p<0.001). The table also indicates that when Chinese evaluators are with Danish users, as well as Danish evaluators with Chinese users, despite the number seen, the important problems are more than minor problems (85 vs. 73; 82 vs. 71), the difference is not significant (p=0.340; p=0.374).

Summary of the Hypothesis 2

Hypothesis 2: Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

- H2 a (in Denmark): When with Danish users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.
- H2 b (in China): When with Chinese users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

Both H2a and H2b were supported by the data. When doing tests with Danish users, there was no significant difference between the severity ranks given by Danish evaluators, whereas Chinese evaluators rated the problems as “minor” or “important” more than “critical.” When doing tests with Chinese users, Chinese evaluators rated the problems as “important” significantly more than “minor” or “critical,” whereas Danish evaluators rated the problems as “minor” or “important” more than “critical.” Hypothesis 2 was confirmed. The findings indicate that Chinese evaluators have the tendency of giving the middle rank to problems when they are asked to rate the severity, especially when doing tests with Chinese users. Danish evaluators tend to give a similar number of minor, important and critical ranks to the problems. However, when doing tests with foreign users, both Danish and Chinese evaluators are not willing to rate the problems as “critical.”

In sum, Section 4.2 has compared the usability problems identified by evaluators in different cultural settings. As discussed in section 3.1.1, the users’ agreement on the evaluators’ problem findings and severity ratings is important in examining the evaluators' understanding of the users' problems. "Usability is ultimately determined by the end user, not an expert evaluator, so realness of problems needs to be established by the user." (Hartson et al. 2001, p. 387) For example, section 4.2.2 shows there is a significant difference on the evaluators' severity ratings, but we cannot tell whether the severity ranks given by local evaluators or foreign evaluators are more "correct." The next section analyzes the consistency of the evaluators' and users' problems in order to examine whether local evaluators understand users better than do foreign evaluators.

4.3 Comparison of the Evaluators' and Users' Usability Problems in Different Cultural Settings

4.3.1 Comparison of the Usability Problems Found by Evaluators and Users in Different Cultural Settings (Hypothesis 3)

Hypothesis 3: The usability problems found by the local evaluators are more consistent with the usability problems found by the target users, compared to the usability problems found by the foreign evaluators and target users:

- H3 a (in Denmark): The usability problems found by Danish evaluators are more consistent with the usability problems found by Danish users, compared to the usability problems found by Chinese evaluators and Danish users;
- H3 b (in China): The usability problems found by Chinese evaluators are more consistent with the usability problems found by Chinese users, compared to the usability problems found by Danish evaluators and Chinese users.

In order to examine Hypothesis 3, the usability problems found by users and evaluators are analyzed in this section. As described in section 3.5.2.1, the comparison of the users' and evaluators' problems is based on the original problem list from the pilot study. Considering the users' ability of finding usability problems, they were asked to mark the problems that they experienced in the test. Each problem written down by evaluators after each test was assigned a number that corresponded to the problem number on the list. Table 24 shows the number of usability problems found by evaluators and users based on the original problem list.

Table 24: The number of usability problems found by evaluators and users based on the problem list

		UPs found by the users		
		UPs not found	UPs found	Total
UPs found by evaluators	UPs not found	0	404	404
	UPs found	196	397	593
	Total	196	801	997

Note. UP is for "usability problem".

Regarding the usability problems on the original problem list (Appendix 4), users and evaluators found 997 problems totally. Among them, users found 801 problems, whereas evaluators found 593 problems, indicating that there were 397 (801+593-997) problems found by both users and evaluators. 404 problems were experienced only by users, and 196 problems were found only by evaluators. Figure 20 shows that for the problems found by evaluators, only 20% of them were not found by users and 80% problems were also experienced by users. Regarding the problems experienced by users, 59% were identified by evaluators and 41% were not found by evaluators.

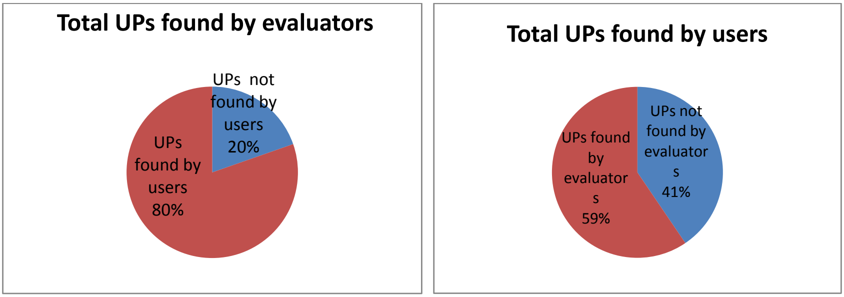


Figure 20: The percentage of problems found by users within the problems found by evaluators, and the percentage of problems found by evaluators within the problems found by the users.

Note. UP is for "usability problem".

Users experienced more problems than the problems found by evaluators. Table 25 shows whether this is the trend for users and evaluators in all the cultural settings or only in some specific cultural settings.

Table 25: The number and rate of the problems found by evaluators and users in different groups

Evaluator-user pairs	UPs found by users but not evaluators	UPs found by evaluators but not users	UPs found by users totally	UPs found by evaluators totally	Total UPs	Rate of usability problems found by users over all UPs	Rate of usability problems found by evaluators over all UPs
DEDU	81	39	162	120	201	0.8060	0.5970
CEDU	68	63	154	149	217	0.7097	0.6866
CECU	146	50	262	166	312	0.8397	0.5321
DECU	109	44	223	158	267	0.8352	0.5918
Totally	404	196	801	593	997	0.8034	0.5948

Note. DEDU means Danish evaluator-Danish user pairs; CEDU means Chinese evaluator-Danish user pairs; CECU means Chinese evaluator-Chinese user pairs; DECU means Danish evaluator-Chinese user pairs; E means evaluators; U means users; UP means usability problem.

Table 25 shows the number and rate of the problems found by evaluators and users in different groups. From the rates of the problems found by the users/evaluators over the problems found by both evaluators and users, we can see that in three conditions (DEDU, CECU, DECU), users found more problems than did evaluators. Only Danish users when with Chinese evaluators found a similar number of problems as did evaluators. But this does not mean that the Danish user-foreign evaluator pairs agreed more than did the others. From Table 25 it is evident that among 149 problems found by evaluators, only 86 (149-63) were also found by users. The user's agreement on the evaluator's problems finding is 57.7%. Figure 21 shows the percentage of problems found by only users, evaluators and the overlapping problems in different conditions.

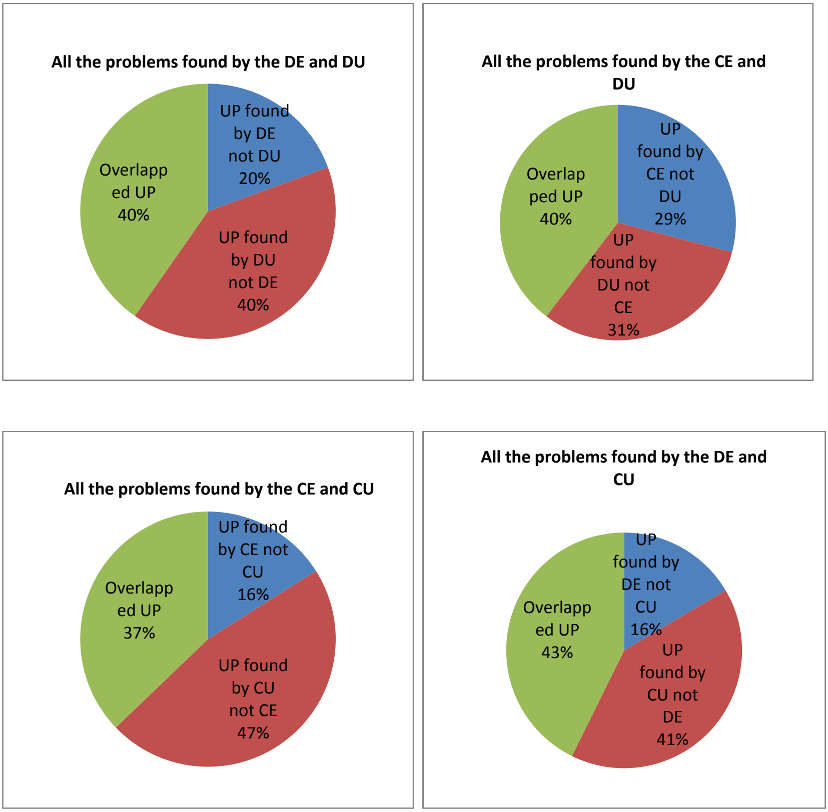


Figure 21: The percentage of usability problems found by only the users, evaluators and the overlapping problems

Note. D for Danish, C for Chinese, E for evaluators, U for users, UP for usability problems.

From Figure 21, we observe that the percentages of the overlapping problems were similar in different cultural settings (37%-43%). All three groups (DEDU, CECU and DECU) had a similar trend, namely, the problems found only by evaluators were not more than 20% (20%, 16%, 16%) and much fewer than the problems found only by users (40%, 47%, 41%). However, the percentage of problems found only by Chinese evaluators when with Danish users was more than that of the other pairs (29%) and similar to the percentage of the problems found only by Danish users (31%).

In order to see whether the problems that evaluators found were really experienced by users, the users' agreement on the problems found by evaluators was calculated. Table 26 shows the users' agreement on the usability problems found by evaluators.

Table 26: The users' agreement on the usability problems found by evaluators

Evaluator - user pairs	UPs found by evaluators totally	UPs found by evaluators but not users	UPs experienced by users within the usability problems that were found by evaluators	Rate of agreement	Chi-Square	Sig.
DE-DU	120	39	81	67.50%	14.700	0.000
CE-DU	149	63	86	57.72%	3.550	0.060
CE-CU	166	50	116	69.88%	26.241	0.000
DE-CU	158	44	114	72.15%	31.013	0.000
Totally	593	196	397	66.95%	68.130	0.000

Note. DE for Danish evaluators, DU for Danish users, CE for Chinese evaluators, CU for Chinese users, UP for usability problem.

For problems found by evaluators, chi-square test was used to examine whether users significantly agreed or not (Hinton et al., 2004, pp. 290-293). Table 26 illustrates that most problems found by evaluators were also experienced by users. The average agreement of the four groups was 66.95%. Danish users significantly agreed with Danish evaluators, and Chinese users significantly agreed with both Chinese and Danish evaluators, whereas the agreement of the Danish users-Chinese evaluators was not significant at the 0.05 level, but marginally significant. When Danish users were with foreign evaluators, the agreement of the evaluators' problem finding was only 57.72%, which was much lower than that of other pairs. On the other hand, Chinese users agreed more with both local and foreign evaluators. The reason may be that Chinese users tended to mark more problems than did Danish users (see Table 27).

Table 27: The discrepancy of the rate of usability problems found by users and evaluators

Evaluator- user pairs	Rate of UPs found by users over all UPs	Rate of UPs found by evaluators over all UPs	Discrepancy
DE-DU	80.60%	59.70%	20.90%
CE-DU	70.97%	68.66%	2.30%
CE-CU	83.97%	53.21%	30.77%

DE-CU	83.52%	59.18%	24.34%
Totally	80.34%	59.48%	20.86%

Note. DE for Danish evaluators, DU for Danish users, CE for Chinese evaluators, CU for Chinese users, UP for usability problem.

Table 27 shows the difference between the rates of the problems found by users and evaluators. Users in all four groups tended to mark more problems than the problems found by evaluators, and the tendency was much clearer for the Chinese users (30.87% and 24.34% vs. 20.9% and 2.30%). If users marked most problems, the chance of agreeing with the evaluators' identified problems was greater.

Summary of Hypothesis 3

Hypothesis 3: The usability problems found by the local evaluators are more consistent with the usability problems found by the target users, compared to the usability problems found by the foreign evaluators and target users:

- H3 a (in Denmark): The usability problems found by Danish evaluators are more consistent with the usability problems found by Danish users, compared to the usability problems found by Chinese evaluators and Danish users;
- H3 b (in China): The usability problems found by Chinese evaluators are more consistent with the usability problems found by Chinese users, compared to the usability problems found by Danish evaluators and Chinese users.

H3a was not entirely supported. 67.50% of usability problems found by Danish evaluators were regarded as problems by Danish users. 57.72% of problems found by Chinese evaluators were agreed upon by Danish users. 67.50% is larger than 57.72%; however, the Chi-square tests show that Danish users significantly agreed on the problems found by Danish evaluators, and the agreement on the Chinese evaluators' problems was also marginally significant (significant at 0.06 level). Thus, even though when doing tests with Danish users, the problems found by local pairs seemed to be more consistent than with the distant pairs, the local pairs and distant pairs had the same tendency in the problem agreement.

H3b was not supported either. Chinese users significantly agreed with the problems found by both local and foreign evaluators (69.88%, 72.15%), and thus Hypothesis 3 is rejected. Most usability problems found by evaluators were also regarded as problems by users.

4.3.2 Comparison of the Usability Problem Severity Given by Evaluators and Users (Hypothesis 4)

Hypothesis 4: The usability problem severity ranks given by local evaluators and target users are more consistent than the problem severity ranks given by the foreign evaluator and target users.

- H4 a (in Denmark): The usability problem severity ranks given by Danish evaluators and Danish users are more consistent than the problem severity ranks given by Chinese evaluator and Danish users;
- H4 b (in China): The usability problem severity ranks given by Chinese evaluators and Chinese users are more consistent than the problem severity ranks given by the Danish evaluator and Chinese users.

As introduced in section 3.5.2.2, Spearman correlation was used to examine the evaluators' and users' severity ratings (see Table 28).

Table 28: The Spearman correlation between evaluators' and users' severity ranks

Evaluator- user pairs	N	Spearman correlation coefficient	Sig. (2-tailed)
Totally	397	0.200^{**}	0.000
DE-DU	81	0.297^{**}	0.007
CE-DU	86	0.278^{**}	0.009
CE-CU	116	0.176	0.058
DE-CU	114	0.109	0.250

Note. **Correlation is significant at the 0.01 level (2-tailed), DE for Danish evaluators, DU for Danish users, CE for Chinese evaluators, CU for Chinese users.

Table 28 shows the Spearman correlation between evaluators' and users' severity ratings for the problems found by both evaluators and users. The results show that the correlation between all evaluators' and users' severity rating was significant (r= 0.200, p<0.001). This indicates that users generally agreed with the severity that the evaluators gave.

The separate analyses of the Danish pairs, Chinese evaluator-Danish user pairs, Chinese pairs, and Danish evaluator-Chinese pairs indicate that:

- Danish pairs' severity ranks were significantly correlated (r_{DEDU}=0.297 p_{DEDU}=0.007).

- The correlation of Chinese evaluator-Danish user pairs' severity rating was significant (r_{CEDU}=0.278 p_{CEDU}=0.009).
- The correlation of the Chinese pairs' severity rating was marginally significant (r_{CECU}=0.176, p_{CECU}=0.058).
- The Danish evaluator-Chinese user pairs' severity ranks were not significantly correlated (r_{DECU}=-0.109, p_{DECU}=0.250).

Summary of the Hypothesis 4

Hypothesis 4: The usability problem severity ranks given by the local evaluators and target users are more consistent than the problem severity ranks given by the foreign evaluator and target users.

- H4 a (in Denmark): The usability problem severity ranks given by Danish evaluators and Danish users are more consistent than the problem severity ranks given by Chinese evaluator and Danish users;
- H4 b (in China): The usability problem severity ranks given by Chinese evaluators and Chinese users are more consistent than the problem severity ranks given by the Danish evaluator and Chinese users.

H4a was not supported. When with Danish users, the severity ranks given by local pairs and distant pairs had a similar trend. The correlations between both local and distant pairs were significant.

H4b was supported by the data. When with Chinese users, the severity ranks given by Chinese evaluators and users were marginally significantly correlated. However, the correlation of the severity ranks given by Danish evaluators and Chinese users was not significant. Thus, the severity ranks given by Chinese pairs were more consistent than for the Danish evaluator–Chinese user pairs.

Thus, Hypothesis 4 is confirmed in the Chinese settings, but not in Danish settings. In the thinking aloud usability testing settings, it is difficult to give a simple assumption. The assumption depends on the evaluators' and users' specific cultural backgrounds. In the following sections, the evaluators' and users' communication patterns are analyzed to examine how they communicate in order to find the usability problems.

4.4 Communication Analysis

In order to investigate the second sub-research question---"To what extent does the evaluators' and users' cultural background influence the evaluators' and users' communications in the thinking aloud usability testing,"---the evaluators' and users' communication patterns in the test were analyzed. As mentioned in section 2.3.2.2, this is a more explorative study, and thus four themes instead of hypotheses were put forward. In the following sections, the four themes are analyzed, but before analyzing the themes, the selected videos are first discussed.

4.4.1 The Selected Videos in the Communication Analysis

As introduced in section 3.5.3.1, the selected videos used for analyzing the communication patterns were the tasks related to the images. There were 16 evaluators and each evaluator did four tests. The average time for each video was around 12 minutes. Table 29 shows the duration of the selected videos in each condition.

Table 29: Duration of the videos related to the tasks of images[3]

	DEDU	CEDU	CECU	DECU
Duration (S)	8427	14019	12058	12089

Table 29 shows that the duration of the Danish pairs is less than that of the other pairs. One way ANONA in SPSS can be used to examine whether there is significant difference between the duration of the videos related to the task of images. In order to do one way ANONA, there are three assumptions that need to be satisfied (Hinton et al., 2004; Michael, 2009): 1) independency, 2) normally distributed populations, and 3) equal variances. The image-related videos were from different tests which can be seen as independent to each other. One sample Kolmogorov-Smirnov test (K-S test) which compares the scores in the sample to a normally distributed population is used to examine whether the distribution of the data here is a normal distribution or not (Field, 2005p 93-96). The results show the distribution of the sample is not significantly different from the normal distribution ($Z_{K\text{-}S}$=1.146, p=0.144), which indicates the data here correspond to the normal distribution (Hinton et al., 2004, p. 32). The third assumption is the homogeneity of the variances. However, Levene's test which is used to analyze the

[3] Note: For all the tables in section 4.4 and section 4.5, DEDU means Danish evaluator-Danish user pairs; CEDU means Chinese evaluator-Danish user pairs; CECU means Chinese evaluator-Chinese user pairs; DECU means Danish evaluator-Chinese user pairs.

homogeneity of variances shows that the variances are unequal (F=3.854, p=0.014). In this situation, since the other two assumptions have been met, one way ANOVA can still be used. There is a way to deal with the violation of the equal variances, which is the “equal variances not assumed” option in the post hoc multiple comparisons in one way ANOVA. Of course, the nonparametric test of Kruskal-Wallis test could also be used to do the analysis, but it is normally used when the distribution is not normal (Michael, 2009).

Thus for the analysis here, we used the one way ANOVA to compare the difference of the four groups and used the “equal variances not assumed” option in the post hoc multiple comparisons to compare all the pair-wise. The results of one way ANOVA show that there was significant difference between the duration of the four groups’ videos. Tamhane’s T2 in the post hoc analysis shows that the duration of the Danish pairs was significantly less than the duration of Chinese evaluator-Danish user pairs and Chinese evaluator-Chinese user pairs (p=0.016, p=0.017); it was also marginally less than the Danish evaluator-Chinese user pairs’ (p=0.074), but there was no significant difference between the durations of CEDU, CECU and DECU pairs.

Since the duration of the selected videos varied, it was deemed to be better to compare the rate of the time for state event or the rate per minute for the point event in the following analyses, instead of using absolute time.

4.4.2 Theme 1

Theme 1: The time spent on the users’ three state events---thinking aloud, talking rather than thinking aloud, and silence --- is different in different cultural settings.

As discussed in section 2.3.1.2 and section 2.3.2.2, the users may “think aloud,” “talk rather than think aloud” and “keep silence” in the usability testing. Theme 1 is investigated to see which states are most common in the tests and whether there is a difference for the state events in different cultural settings. Table 30 shows the time and rate of time spent on each state event within different groups.

Table 30: The time and the rate of time spent on each state event within each group

	TA (S)	Talk rather than TA (S)	Silence (S)	Total (S)	TA%	Talk%	Silence%
DEDU	1505	6533	221	8259	0.182	0.791	0.027
CEDU	2975	10377	417	13769	0.216	0.754	0.030
CECU	398	9861	1528	11787	0.034	0.836	0.130
DECU	312	9634	1893	11839	0.026	0.814	0.160

Note. TA is for “think aloud”.

Kruskal-Wallis test in SPSS was used to examine whether the time spent on thinking aloud, talking, and silence was significantly different within each group. Here, we could not use One-Way ANOVA, since the test of homogeneity of variances showed that the three state events were from the populations of unequal variances (F_{DEDU}=19.950, p<0.001; F_{CEDU}=15.507, p<0.001; F_{CECU}=14.602, p<0.001; F_{DECU}=12.900, p<0.001); and the Kolmogorov-Smirnov test (Field, 2005; Hinton et al., 2004) shows that the distributions of the three state events in the four groups are significantly different from that of the normal distribution (Z_{DEDU}=1.546, p=0.017; Z_{CEDU}=1.483, p=0.025; Z_{CECU}=1.528, p=0.019; Z_{DECU}=1.630, p=0.010). So the nonparametric analysis Kruskal-Wallis test was used, instead of the One-Way ANOVA. The results show that there are significant differences between the three state events in all groups, see Table 31.

Table 31: Kruskal-Wallis Test of the time spent on the three state events

Pairs	df	Chi-Square	Sig.
DEDU	2	31.507**	.000
CEDU	2	31.528**	.000
CECU	2	35.484**	.000
DECU	2	35.786**	.000

The significant difference shows only that there are differences between the three state events, but does not differentiate which event is significantly different from the others. The Post hoc multiple comparison test is normally used to compare all different combinations of the groups (Coolican, 2004; Field, 2005; Hinton et al., 2004). However, the post hoc multiple comparison test is only in one way ANONA and general linear model (GLM) analysis, not in Kruskal-Wallis analysis in SPSS (Hinton et al., 2004, p. 267). We therefore cannot compare all the combinations of every two groups because the Type 1 error will be increased (Coolican, 2004, p. 488). There is a good way to do the non-parametric post hoc analysis, which is to use multiple Mann-Whitney tests and, at the same time, make an adjustment to ensure that the type 1 error does not build up to more than 0.05 (Field, 2005, p. 550). This method is called “Bonferroni correction” which uses a critical value of 0.05 divided by the number of tests that will be conducted, instead of using 0.05 as the critical value for significance (Field, 2005, p. 550).

In this study for each group, in order to compare the three state events, three Mann-Whitney tests were conducted, for which the critical level of significance is 0.05/3=0.0167. If the calculated significance value is less than 0.0167, the difference will be significant, whereas if the calculated significance value is greater than 0.0167, there will be no significant difference. Table 32 shows the results of the Mann-Whitney tests for each group.

Table 32: Multiple Comparisons of the time spent on the state events

Pairs	State Events Comparison		Mann-Whitney U	Asymp. Sig.
DEDU	TA	Talk	18.000*	.000
	TA	Silence	61.500*	.011
	Talk	Silence	0.000*	.000
CEDU	TA	Talk	21.000*	.000
	TA	Silence	51.500*	.003
	Talk	Silence	0.000*	.000
CECU	TA	Talk	0.000*	.000
	TA	Silence	56.500*	.006
	Talk	Silence	0.000*	.000
DECU	TA	Talk	0.000*	.000
	TA	Silence	32.000*	.000
	Talk	Silence	9.000*	.000

Note. * The difference is significant at the 0.05 level (less than 0.0167 for each pair), TA is for "think aloud".

For all pairs, the time spent on "talk rather than TA" was significantly longer than the time on thinking aloud or silence ($p<0.001$, less than 0.0167). For the tests with Danish users, the time spent on thinking aloud was significantly longer than the time on silence ($p_{DEDU}=0.011$, $p_{CEDU}=$-003, both less than 0.0167). For the tests with Chinese users, the time spent on thinking aloud was significantly less than the time on silence ($p_{CECU}=0.006$, $p_{DECU}<0.001$, both less than 0.0167). The analysis shows that regardless of whether with local or foreign evaluators, all users tended to talk (such as explanation) more than think aloud or keep silent. Further, Danish users tended to think aloud more than keep silent, whereas Chinese users tended to be silent more than think aloud.

From Table 30, it appears that Danish users when with foreign evaluators spent more time on thinking aloud than when with local evaluators, whereas Chinese users when with foreign evaluators spent more time on silence than when with local evaluators. Since the total time of the tests was different, it was better to compare the rate of time, not the absolute time. T test was used to compare the users' state events when they were with local and foreign evaluators (Table 33).

Table 33: T test of the users' state events when with local and foreign evaluators

	Danish users with local and foreign evaluators			Chinese users with local and foreign evaluators		
	TA	Talk	Silence	TA	Talk	Silence
T	-0.434	0.301	-0.201	0.404	0.191	-0.573
df	30	30	30	30	30	30
Sig	0.667	0.765	0.842	0.689	0.850	0.571

Note. TA is for "think aloud".

Table 33 shows that there was no significant difference for users with local and foreign evaluators for all three state events, and for both Danish and Chinese users (p from 0.571 to 0.850). The users spent similar time on thinking aloud/ talking/ silence when with local and foreign evaluators.

The rate of time spent on thinking aloud for Danish users was significantly longer than that for Chinese users' (U=238.000, p<0.001). Danish users spent around 20% of the time on thinking aloud, whereas Chinese users spent only around 3% of the time on it. On the other hand, the rate of time spent on silence for Chinese users was significantly longer than it was for Danish users (U=174.000, p<0.001). Chinese users spent around 14.5% of the time on silence, whereas Danish users spent around 2.55% of the time on silence.

Summary

The findings in theme 1 are:

1) For all pairs, the time spent on "talking rather than thinking aloud" was significantly longer than the time on thinking aloud or silence.
2) For Danish users, the time spent on thinking aloud was significantly longer than the time on silence.
3) For Chinese users, the time spent on thinking aloud was significantly less than the time on silence.
4) For all three state events, there was no significant difference when the users were with local and foreign evaluators.
5) Danish users spent significantly longer time on thinking aloud than did Chinese users.

6) Chinese users spent significantly longer time on silence than did Danish users.

4.4.3 Theme 2

Theme 2: Users' verbal behaviors are different in different cultural settings.

Theme 2 is proposed to examine the users' verbal behaviors. Danish and Chinese users with local and foreign evaluators may behave differently. Table 34 shows four types of verbal behaviors in different cultural settings.

Table 34: Users' point-event behaviors

Point-event behaviors	Result	DEDU	CEDU	CECU	DECU
Negative comments	Total number	167	215	153	165
	Total rate per minute	18.84	14.03	12.37	11.67
Positive comments	Total number	40	95	67	67
	Total rate per minute	4.1	7.99	5.27	5.88
Culture related comments	Total number	2	12	7	20
	Total rate per minute	0.2	1	0.5	1.7
Suggestions	Total number	56	115	94	62
	Total rate per minute	6.84	8.12	7.13	5.25

Table 34 shows the total number of the users' point-event behaviors and the rate of the behavior per minute. The "rate per minute" is calculated by the total number of the behaviors in each test divided by the duration of each test session. For example, if there are 15 negative comments and the test duration is 10 minutes, the rate per minute will be 1.5. The total rate per minute in Table 34 is the sum of the 16 test rates per minute in each group. Because of the different durations of the test sessions, it was better to compare the point-event behaviors' "rate per minute" in each condition instead of the number.

SPSS 16.0 was used to examine the difference of the users' behaviors in different groups. The results show Danish users' negative comments per minute were significantly higher than those of Chinese users' (t=2.230, p=0.029), but there was no significant difference between Danish users when they were with Danish and Chinese evaluators (t=1.604, p=0.119). There was also no significant difference between Chinese users when with local and foreign evaluators (t=0.276, p=0.784).

For the positive comments rate per minute, there was no significant difference between Danish and Chinese users (t=0.319, p=0.750). Further analysis of the positive comments per minute with Danish users when with local and foreign evaluators showed Danish users' positive comments had no significant difference (t_D=1.610, p=0.118), even though from Table 34 it appears Danish users gave many more positive comments when with local evaluators than when with foreign evaluators. There was also no significant difference for Chinese users when with local and foreign evaluators (t_c=0.379, p=0.708).

Table 34 shows that for both Danish and Chinese users, when with foreign evaluators, they gave more culture related comments than when they were with local evaluators. In order to see whether the trend was significant or not, we could not use T test since Levene's test, after examining the equality of variances, showed that the two groups did not have equal variances (F=15.946, p<0.001) (Hinton et al., 2004, pp. 111-112). T test is a parametric test which requires that the data from the two samples comes from "populations with equal variances (the homogeneity of variance assumption)" (Hinton et al., 2004, p. 107). Thus, we used Mann-Whitney U test (Hinton et al., 2004, pp. 124-129), the results of which show that there was significant difference between the users' culture related comments per minute when with local and foreign evaluators (U=330.500, p=0.005). Danish users when with foreign evaluators gave more culture related comments than when with local evaluators, and it was marginally significant (U_D=87.500, p=0.054). Chinese users gave significantly more culture related comments to foreign evaluators than they did to local evaluators (U_C=76.500, p=0.032). There was no significant difference between the culture related comments given by Danish and Chinese users (U=426.000, p=0.178).

For the suggestions per minute, there was no significant difference between Danish and Chinese users' (t=1.191, p=0.238). There was also no significant difference between Danish users or Chinese users when with local and foreign evaluators (t_D=0.797, p=0.432; t_C=1.299, p=0.204).

Summary

The findings in theme 2 are:

1) Danish users gave significantly more negative comments than did Chinese users.
2) Users gave more culture related comments to foreign evaluators than they did to local evaluators. The tendency was especially stronger for the Chinese users.
3) There was no significant difference for the positive comments and suggestions given by users in different cultural settings.

4.4.4 Theme 3

Theme 3: Evaluators' verbal behaviors are different in different cultural settings.

In theme 3, the evaluators' verbal behaviors are analyzed. Table 35 shows the evaluators' point-event behaviors in different cultural settings.

Table 35: Evaluators' point-event behaviors

Point event behaviors	Results	DEDU	CEDU	CECU	DECU
Affirmative	Total number	483	833	556	651
	Total Rate per minute	56.54	61.19	43.91	53.85
Directed probing	Total number	36	26	43	52
	Total Rate per minute	3.89	2.11	3.61	3.88
Digging deeper probing	Total number	170	147	170	202
	Total Rate per minute	19.52	11.39	13.7	16.23
Using question as the reminder to think aloud	Total number	9	5	15	19
	Total Rate per minute	1.09	0.28	1.35	1.45
Classic reminder	Total number	1	0	1	1
	Total Rate per minute	0.08	0	0.09	0.08
Help behavior	Total number	5	12	20	12
	Total Rate per minute	0.56	0.8	1.53	0.8
No clear answer or help, but encourage	Total number	2	6	6	9
	Total Rate per minute	0.35	0.32	0.38	0.64

For the affirmative responses shown in Table 35, it appears that foreign evaluators tended to give more affirmative responses to show their active listening than did local evaluators (61.19 vs. 56.54; 53.85 vs. 43.91), but the trend was not significant (t=1.556, p=0.125). Examining the tendency with Chinese users, the foreign evaluators tended to give more affirmative responses than did local evaluators, and it was marginally significant (t=0.542, p=0.090).

From Table 35, we can see that Danish evaluators had more directed probing and digging deeper probing behaviors. The difference of the Danish and Chinese evaluators' directed probing behaviors was not significant (t=1.217, p=0.228). When with Danish users, even though the Danish evaluators' directed probing behaviors were more than those of Chinese evaluators' (3.89 vs. 2.11), it was not significant (t=1.370, p=0.181).

The Danish evaluators had significantly more digging deeper probing behaviors than did Chinese evaluators (t=2.361, p=0.021). When with Chinese users, the Danish and Chinese evaluators' digging deeper probing behaviors were not significantly different (t=0.941, p=0.354), whereas when with Danish users, the Danish evaluators probed significantly more questions to dig deeper or ask the users to clarify (t=2.230, p=0.033) than did Chinese evaluators with Danish users.

When with Chinese users, the evaluators tended to use more reminders than when with Danish users, see Table 35 (16 and 20 vs. 10 and 5) and it was marginally significant (t=1.797, p=0.077). When with Danish users, Danish evaluators used more reminders to keep the users talking than did Chinese evaluators (10 vs. 5), and it was marginally significant (U=92.000, p=0.074). Normally, the evaluators used questions to ask the users to keep talking (such as, "what are you thinking now?") instead of the classic reminder which was "please keep talking" or "please speak out your thoughts." In all 64 tests, there were only 3 classic reminders, compared with 48 reminders in the form of questions (U=1350.500, p<0.001).

The help behavior means that the evaluators were willing to help the user to figure out some problems, and the "no clear answer or help" means that the evaluators did not want to give solutions but encouraged the users to figure out the problems by themselves. Table 36 gives both data of the "help behavior--A" and the "no help but encourage--B." (A+B) refers to the occasion that needs the evaluators' help. Thus, B/(A+B) signifies the rate of not willing to help, and A/(A+B) signifies the rate of willing to help. Table 36 shows the rates. Chinese evaluators seemed to be more willing to provide help than were Danish evaluators (0.71 and 0.8 vs. 0.62 and 0.56).

Table 36: The rates of the evaluators' helping behavior and not helping behavior

Helping behaviors	Results	DEDU	CEDU	CECU	DECU
Help behavior (A)	Total Rate per minute	0.56	0.8	1.53	0.8
No clear answer or help, but encourage (B)	Total Rate per minute	0.35	0.32	0.38	0.64
Total (A+B)	Total Rate per minute	0.91	1.12	1.91	1.44
Rate of help behavior	A/ (A+B)	0.62	0.71	0.80	0.56
Rate of not willing to help	B/ (A+B)	0.38	0.29	0.20	0.44

Summary

The findings in theme 3 are:

1) When with Chinese users, foreign evaluators gave more affirmative responses than did local evaluators, and it was marginally significant.
2) There was no significant difference in giving directed probes in all cultural settings.

3) The Danish evaluators gave significantly more digging deeper probes than did Chinese evaluators. The tendency was especially significant when with Danish users.
4) For the reminders, the evaluators used significantly more non-classic reminders than classic reminders to keep the users talking. Evaluators gave more reminders to Chinese users than to Danish users. When with Danish users, Danish evaluators gave more reminders than did Chinese evaluators.
5) Chinese evaluators seemed to be more willing to provide help than were Danish evaluators.

4.4.5 Theme 4

Theme 4: The way to communicate usability problems is different in different cultural settings.

The above analysis was based on the selected test sessions. From the above analysis, we can see how the users and evaluators communicated in a test session in order to find the positive and negative aspects of the culturally localized application. Since the main task for the evaluator was to find usability problems through listening to the users' talk and observing the users' behavior, it is useful to analyze the whole session in order to see how the evaluators facilitated the test and communicated with the users. Some behaviors and communication may be directly related to problems, whereas others may not be related to the problems but to some related discussions of the application, such as an explanation about the culture. The above analysis is about the users' and evaluators' communication patterns in the test. Theme 4 focuses on the analysis of communication patterns for usability problems.

In this study, the problems were coded as UPUs and UPEs (section 2.3.2.2 and section 3.5.3.2). As introduced in section 2.3.2.2, the UPU and UPE were coded based on the consideration of the evaluators' roles in finding the problems. For the UPU, the evaluator's main role is the receiver, who gets the problem by observing the user's task performing and listening to the user's "thinking aloud" or comments. For the UPE, the evaluator's main role is the sender, who has the problem in mind and needs to get the confirmation from the users. The coders coded UPU and UPE based on the observation of the video (see section 3.5.3.2 and Appendix 14).

As discussed in section 2.3.2.2, theme 4 includes four sub-themes: theme 4a to Theme 4d. The sub-themes are analyzed based on the UPU and UPE. Before analyzing the theme 4a to theme 4d, we need to make sure that the UPUs and UPEs are identified in different ways, and thus it is necessary for us to divide the usability problems into UPUs and UPEs. Lag sequential analysis is used to analyze the starts of the problems. Lag sequential analysis in Observer is based on the associations between coded events (Eide et al., 2003) and is used to analyze the order of the events. For example, from the lag sequential analysis, we can see which behavior follows from negative comments. In this study, we focus mainly on analyzing the behavior of the start of the problem. The problem was coded as state event with both start and end. As explained earlier, the start was the user's verbal or task performing behavior or the evaluator's verbal behavior, whichever was relevant to the problems written down by the evaluators. The behavior before the start of the problem was another topic. The start of the problem could be seen as a sign showing that there was a problem or as a way to elicit the problem. From the lag sequential analysis, we can see the behaviors of starting the UPU and UPE, see Table 37 and Table 38.

Table 37: The number of the behaviors for starting the UPU

Users' and evaluators' Behaviors	DEDU	CEDU	CECU	DECU
Classic thinking aloud	4	0	0	1
Silence	1	0	0	0
Negative comments	22	43	20	27
Positive comments	0	2	0	1
Suggestions	10	19	6	8
Culture related comments	0	1	0	1
Asking questions by the user	4	4	3	2
Directed probing	1	2	3	2
Digging deeper probing	4	1	2	7
Using question as the reminder to think aloud	0	0	0	1
Probing: other (general questions)	10	12	14	9
Help behavior	0	0	0	1

Table 37 shows the behaviors of starting the UPUs. We can see that most UPUs started from the users' negative comments and suggestions. When users gave negative comments or suggestions, it could indicate a potential problem. Sometimes the problems also started from the users' questions. When with foreign evaluators, the users could start with positive comments or culture related comments in order to talk about a problem, even though it happened infrequently. Users' classic thinking aloud or silence could also show a problem. Evaluators' general questions, such as "What do you think the images provided here?" played a more important role in eliciting the users' comments relevant to UPUs, compared to the other probing behaviors, which indicated many problems that could be from interviewing the users.

Table 38 shows the number of the behaviors for starting a UPE. Most UPEs started from the evaluators' directed probing behaviors.

Table 38: The number of the behaviors for starting the UPE

Users' and evaluators' Behaviors	DEDU	CEDU	CECU	DECU
Negative comments	0	0	0	1
Suggestions	1	0	0	1
Short reply	0	0	0	1
Directed probing	8	5	3	11
Digging dipper probing or ask the user to clarify	1	0	0	2
Probing: other (general questions)	1	1	0	0
Help behavior	0	0	1	1

The lag sequential analysis showed that most UPUs started from the users' negative comments and suggestions, whereas most UPEs started from the evaluators' directed probing behaviors. The analysis indicates that UPUs and UPEs do involve different communication patterns, which should be analyzed separately.

Therefore, in theme 4, the communication patterns for the UPUs and UPEs are analyzed separately. In order to analyze the behaviors within the period of the UPUs and UPEs, the data should be selected first by using Observer 8.0. For example, when analyzing the behaviors in the UPU period, the data for the UPEs and the data unrelated to problems will be excluded. In the following sub-sections, theme 4a to theme 4d are analyzed.

4.4.5.1 Theme 4a

Theme 4a: The number and duration of the UPU and UPE may be different in different cultural settings.

Theme 4a is developed in order to investigate the number and duration of the UPU and UPE in different cultural settings. Table 39 shows the number and the mean duration of the UPUs and UPEs in this study.

Table 39: The number and the mean duration of UPUs and UPEs

	Total UPE number	UPE mean duration (Second)	Total UPU number	UPU mean duration (Second)	Total UP	Rate of UPE	Rate of UPU
DEDU	11	37.81	67	52.28	78	0.141	0.859
CEDU	6	40.83	95	44.74	101	0.059	0.941
CECU	5	70.60	61	62.61	66	0.076	0.924
DECU	19	44.42	72	54.32	91	0.209	0.791
Total	41	45.31	295	52.49	336	0.122	0.878

Note. UP is for "usability problem".

From Table 39 we can see that the number of UPUs was larger than the number of UPEs in all cultural settings, which implies that most problems were from the users' task performing behaviors or speech. But in each cultural setting, there were still some UPEs, which indicate that some problems were not from the users' behaviors, but from the evaluators' directed probing. Mann-Whitney U test was used to examine whether the number of UPU was significantly larger than the number of UPE. The results show that the number of UPE was significantly smaller than the number of UPU ($U<0.001$, $p=0.021$), which indicates that most problems were from the users' behaviors, not the evaluators' mind.

But in each cultural setting, there were UPEs, which showed that the evaluators did not always follow the user's current thoughts and passively listen, but also used directed probes to call the user's attention to the issues that they were interested in. From Table 39, we can see that the rates of Danish evaluators' UPEs were two or three times larger than those of Chinese evaluators' rates (0.141 and 0.209 vs. 0.059 and 0.076), which implies that Danish evaluators were more active than were Chinese evaluators, indicating that there were more UPEs.

For the durations of the UPU and UPE, there was no significant difference ($t_{(334)}=0.963$, $p=0.336$). Further analysis of the durations of the UPE and UPU in each cultural setting showed that there was no significant difference for the duration of UPU and UPE in all the cultural settings ($t_{DEDU(76)}=0.810$, $p=0.420$; $t_{CEDU(99)}=0.296$, $p=0.768$; $t_{CECU(64)}=0.353$, $p=0.725$; $t_{DECU(89)}=0.872$, $p=0.386$). The results imply that regardless of whether the problem was identified from the users' speech and behavior or from the evaluators' probing behavior, the time spent on discussing the problems was similar.

Summary

The findings in theme 4a are:

1) UPUs were significantly more than UPEs in all the cultural settings, implying that the problems were mainly from the users' behaviors, not the evaluators'.
2) The rates of Danish evaluators' UPEs were two or three times larger than the rates of the Chinese evaluators' UPEs.
3) There was no significant difference for the duration of UPUs and UPEs in all cultural settings. The users and evaluators spent similar time on discussing the UPUs and UPEs.

4.4.5.2 Theme 4b

Theme 4b: The time spent on discussing UPU/ UPE with different severity is different in the four cultural settings.

One of the objective measures of the severity is the time spent on the problem (Hassenzahl, 2000). The evaluators and users are supposed to spend more time on dealing with or discussing the critical or important problems than on minor problems. The severity of the problem was decided by the evaluators after each test. Whether the evaluators in different cultural settings considered the time spent on the problem when they were rating its severity was examined. Table 40 shows the time spent on the UPUs and UPEs with different severities in different cultural settings.

Table 40: The time spent on minor, important and critical problems for each cultural setting

UP Types	UPE				**UPU**			
Pairs	DEDU	CEDU	CECU	DECU	DEDU	**CEDU**	CECU	**DECU**
N_{Minor}	3	3	1	7	23	**20**	13	**14**
Total Duration(S)	153	118	33	552	997	**767**	746	**490**
Mean Duration(S)	51.00	39.33	33.00	78.86	43.35	**38.35**	57.38	**35.00**
$N_{Important}$	3	2	3	10	19	**42**	38	**30**
Total Duration(S)	110	66	313	262	1184	**1799**	2577	**1651**
Mean Duration(S)	36.67	33.00	104.33	26.20	62.32	**42.83**	67.82	**55.03**
$N_{Critical}$	5	1	1	2	25	**33**	10	**28**
Total Duration(S)	153	61	7	30	1322	**1684**	496	**1770**
Mean Duration(S)	30.60	61.00	7.00	15.00	52.88	**51.03**	49.60	**63.21**

Note. UP is for "usability problem."

One-Way ANOVA in SPSS was used to examine whether the time spent on critical, important and minor problem varies. The results show that there was no significant correlation between the time spent on the problems and their severities for all 336 problems in the image-related tasks ($F_{(2, 333)}$=0.952, p=0.387). For the 41 UPEs in all four groups, there was no relation between the time spent on discussing the problems and the severity of the problems ($F_{(2, 38)}$=1.319, p=0.279). For all 295 UPUs, the time spent on minor, important and critical problems was marginally different ($F_{(2, 292)}$=2.240, p=0.108). Further multiple comparisons analysis (LSD) shows that the time spent on minor problems was significantly less than the time on important problems (p=0.045) and marginally less than on critical problems (p=0.080).

The time spent on UPEs with different severities in each cultural setting was not analyzed, since in some groups there were only one or two problems, for which it was difficult to do statistical analysis. But from the observation of Table 40, we can see that there was no consistent tendency for the time spent on UPEs with different severities.

For the UPUs in the local conditions (DEDU, CECU), there was no significant difference between the time spent on the problems with the three severity ranks ($F_{(2,125)}$=1.455, p=0.237). For the UPUs in foreign conditions, the time spent on the problems with different ranks was different ($F_{(2,164)}$=3.594, p=0.030). Further multiple comparisons analysis (LSD) shows that the time spent on critical problems was significantly longer than the time on minor problems (p=0.009).

Regarding the UPUs in the Danish evaluator-Chinese user pairs, the time spent on critical, important and minor problems was marginally different ($F_{(2, 69)}$ =2.651, p=0.078). Further multiple comparisons analysis (Tamhane's T2) shows that the time spent on minor problems was significantly smaller than the time on critical problems (p=0.038) and marginally smaller than the time on important problems (p=0.096). This implies that when doing tests in another culture, the evaluators tended to spend more time on discussing critical and important problems rather than on minor problems if the problem was from the users' behavior or comments. From Table 40 we can see that Chinese evaluator-Danish user pairs also had this trend, although it was not significant ($F_{(2,92)}$=1.125, p=0.329).

Summary

The findings in theme 4b are:

1) For the UPEs, there was no tendency of spending more time on critical or important problems than on minor problems.
2) For the UPUs, the evaluators spent less time on minor problems than on important or critical problems. However, for the UPUs found by local evaluators, there was no significant difference for the time spent on problems with different severities. For the distant pairs' UPUs, foreign evaluators spent significantly more time on critical problems than on minor problems. The tendency was especially significant for the Danish evaluator-Chinese user pairs.

4.4.5.3 Theme 4c

Theme 4c: Users' communication patterns for the UPU and UPE are different in different cultural settings.

The theme 4c is intended to examine the users' communication patterns in finding UPUs and UPEs. Since UPUs and UPEs involve two different problem information transmitting procedures, during the period of discussing UPUs and UPEs, users may have different communication patterns, and the communication patterns may be affected by the users' and evaluators' cultural backgrounds. In the following section, the users' state-event and point-event behaviors are analyzed within the UPU and UPE periods.

Analysis of the users' state-event and point-event behaviors in the UPU period

1) Analysis of the three state events in the UPU period

Table 41 shows the time and the rate of the time on thinking aloud, talking rather than thinking aloud, and silence.

Table 41: The time and the rate of time on each state event for finding UPUs

Pairs	TA (S)	Talk rather than TA (S)	Silence (S)	Total (S)	TA%	Talk%	Silence%
DEDU	649	2734	104	3487	0.19	0.78	0.03
CEDU	611	3593	48	4252	0.14	0.85	0.01
CECU	115	3306	383	3804	0.03	0.87	0.10
DECU	84	3087	752	3923	0.02	0.79	0.19

Note. TA is for "think aloud".

Table 41 shows the time spent on each state event and also the percentage of time on the event in each cultural setting. We can see that the state event of "talk rather than TA" is longer than the other two in all four groups. In order to examine the difference of the state events in each group, Kruskal-Wallis test is used. Here we cannot use One-Way ANOVA, since one sample Kolmogorov-Smirnov test shows the distribution is significantly different from the normal distribution for all groups (Z_{DEDU}=1.655, p=0.008; Z_{CEDU}=1.659, p=0.008; Z_{CECU}=1.811, p=0.003; Z_{DECU}=1.823, p=0.003). The results of the Kruskal-Wallis test are shown in Table 42.

Table 42: Kruskal-Wallis Test of the time on the three state events

Pairs	df	Chi-Square	Sig.
DEDU	2	28.483**	.000
CEDU	2	33.987**	.000
CECU	2	32.358**	.000
DECU	2	26.582**	.000

Kruskal-Wallis Test results show that for all pairs, the time on the three state events is significantly different (p<0.001) in the UPU period. Further multiple comparison analysis could be conducted to compare every two pairs to see where the significant difference was from. However, as mentioned earlier, the Kruskal-Wallis analysis in SPSS does not provide such function, and thus we used Mann-Whitney tests instead. The critical value of significance in this analysis is 0.05/3=0.0167. See Table 43.

Table 43: Multiple comparisons of the time on the state events

Pairs	State Events Comparison		Mann-Whitney U	Asymp. Sig.
DEDU	TA	Talk	25.000*	0.000
	TA	Silence	78.000	0.061
	Talk	Silence	4.000*	0.000
CEDU	TA	Talk	11.000*	0.000
	TA	Silence	67.000	0.021
	Talk	Silence	0.000*	0.000
CECU	TA	Talk	4.000*	0.000
	TA	Silence	82.000	0.086
	Talk	Silence	11.500*	0.000
DECU	TA	Talk	6.000*	0.000
	TA	Silence	75.000	0.047

	Talk	Silence	34.000*	0.000

Note. * The difference is significant at the 0.05 level (less than 0.0167 for each pair), TA is for “think aloud”.

Table 43 shows that the time spent on “talk rather than TA” is significantly longer than the time on “TA” and “silence” in all the cultural settings, and there is no significant difference between the time spent on “TA” and “silence.” The results indicate that in the UPU period, the users normally “talk” instead of thinking aloud or being silence.

Table 41 indicates that in the UPU period, even though for all the pairs there was no significant difference between the time spent on “TA” and “silence,” it seems that Danish users had longer periods of time on thinking aloud than did Chinese users, and that Chinese users had longer periods of time on being silent than did Danish users. In order to compare the difference between the different cultural settings, it was better to use the rate of the time spent on state event instead of the absolute time because of the difference of the total time. Since Levene’s test for equality of variances shows that only the data of “talk rather than TA” has equal variances ($F=0.152$, $p=0.698$), and the “TA” and “silence” do not have equal variances ($F=20.459$, $p<0.001$; $F=9.820$, $p<0.003$). Since we could not use T test to compare the Danish and Chinese users’ time spent on thinking aloud and silence, we used the nonparametric test Mann-Whitney U test (Hinton et al., 2004). The results show that Danish users took significantly more time on thinking aloud than did Chinese users ($U=823.000$, $p=0.001$), and Chinese users took significantly more time on silence than Danish users ($U=323.000$, $p=0.004$). There was no significant difference between the Danish users’ and Chinese users’ time spent on “talk rather than TA” ($t=0.115$, $p=0.909$).

For both Danish and Chinese users, there was no significant difference between the time spent on thinking aloud, “talking rather than thinking aloud” and silence when with local and foreign evaluators ($U_{DTA}=122.500$, $p_{DTA}=107.500$; $t_{Dtalk}=0.382$, $p_{Dtalk}=0.705$; $U_{Dsilence}=107.500$; $p_{Dsilence}=0.310$; $U_{CTA}=116.000$, $p_{CTA}=0.507$; $t_{Ctalk}=0.469$, $p_{Ctalk}=0.642$; $U_{Csilence}=111.000$; $p_{Csilence}=0.503$).

2) *Analysis of the users’ point events in the UPU period*

Table 44 shows the users’ point events in the UPU period

Table 44: Users' point-event behaviors in the UPU period

Point-event behaviors	Result	DEDU	CEDU	CECU	DECU
Negative comments	Total number	97	119	75	88
	Total Rate per minute	31.47	25.67	19.71	26.53
	Average Rate per minute	1.97	1.60	1.23	1.66
Positive comments	Total number	12	11	6	11
	Total Rate per minute	2.4	2.44	1.28	3.08
	Average Rate per minute	0.15	0.15	0.08	0.19
Culture related comments	Total number	0	4	4	8
	Total Rate per minute	0	1.25	1.77	1.56
	Average Rate per minute	0.00	0.08	0.11	0.10
Suggestions	Total number	29	67	35	26
	Total Rate per minute	8.65	17.13	8.06	12.15
	Average Rate per minute	0.54	1.07	0.50	0.76
Ask questions	Total number	11	11	10	8
	Total Rate per minute	2.63	3.57	2.29	2.96
	Average Rate per minute	0.16	0.22	0.14	0.19
Short reply by the user	Total number	68	49	61	51
	Total Rate per minute	18.79	12.04	13.61	11.30
	Average Rate per minute	1.17	0.75	0.85	0.71

From Table 44 we can see that in the UPU period, there were many negative comments. Every minute there was more than one negative comment for each test (1.23 to 1.97). Many suggestions were also given by users. Every one or two minutes, there was around one suggestion (0.50 to 1.07) for each test.

Table 34: Users' point-event behaviors in the test

Point-event behaviors	Result	DEDU	CEDU	CECU	DECU
Negative comments	Total number	167	215	153	165
	Total rate per minute	18.84	14.03	12.37	11.67
Positive comments	Total number	40	95	67	67
	Total rate per minute	4.1	7.99	5.27	5.88
Culture related comments	Total number	2	12	7	20
	Total rate per minute	0.2	1	0.5	1.7
Suggestions	Total number	56	115	94	62
	Total rate per minute	6.84	8.12	7.13	5.25

Comparing the total rate per minute of Tables 34 and 44, the total rates per minute of both the negative comments and suggestions in the UPU period were more than those in the selected videos. Mann-Whitney test shows that the total rate per minute of the negative comments in the UPU period was significantly more than that in the selected videos (U<0.001, p=0.021). The total rate per minute of suggestions in the UPU period was also significantly more than that in the selected videos (U=1.000, p=0.043). The results indicate that in the UPU period, the negative comments and suggestions in every minute happened more often than those in the test period. Thus, when users have problems, they may give negative comments or suggestions more often.

In the UPU period, the asking question behaviors, positive comments and culture related comments were also analyzed. The culture related comments can be used to discuss the users' feelings of the application. Since it was a wedding invitation application, sometimes the user wanted to talk about wedding issues in their culture in order to make the evaluator understand their comments or suggestions. The cultural issue was not an important issue for understanding the problems, with the result that users in all cultural settings mentioned it infrequently. In the UPU period, there were also positive comments, although this was also not often. Mann-Whitney test results show that the positive comments happened much less frequently in the UPU period than they did in the test (U<0.001, p=0.021). Checking the video, we found that in the UPU period, sometimes the user gave positive comments to compare the positive part with the negative part. For example, the user said, "I don't like this heart, it is too colourful. I like the simple one, such as that heart, only black and white. I think the application could provide more simple images." The whole sentence was in the UPU period, but with some positive comments. Moreover, sometimes questions could also show a problem. When the user did not know how to figure out an issue, he/she sometimes asked questions, which could act as an indicator of a problem. Table 44 also shows that "short reply by the user" happened frequently. "Short reply" usually shows the users' responses to the evaluators' verbal behavior. When the evaluator probed questions, the user could use a short reply such as "yes" to express his/her agreement.

For all the above point event behaviors, there was no significant difference between the local and distant pairs, and there was also no significant difference between the Danish and Chinese pairs. This implied that in the UPU period, the Danish and Chinese users had similar behavior patterns, regardless of whether they were with local evaluators or foreign evaluators. The users' behaviors in finding the UPEs are analyzed next.

Analysis of the users' state-event and point-event behaviors in the UPE period

The users' communication patterns in the UPE period are analyzed in this section.

Table 45: The time and the rate of time spent on each state event for finding UPEs

Pairs	TA (S)	Talk rather than TA (S)	Silence (S)	Total (S)	TA%	Talk%	Silence%
DEDU	24	382	10	416	0.06	0.92	0.02
CEDU	25	220	0	245	0.10	0.90	0.00
CECU	0	295	47	342	0.00	0.86	0.14
DECU	28	690	127	845	0.03	0.82	0.15

Note. TA is for “think aloud”.

Table 45 shows the same trend as the state events in UPUs. Since there were too few tests with the data of the state events, it was impossible to do statistic analysis for each cultural setting. Table 46 shows the number of the tests with each state event.

Table 46: The number of the tests with each state event

Pairs	TA	Talk rather than TA	Silence	Total tests’ number
DEDU	1	6	1	6
CEDU	1	4	0	4
CECU	0	3	2	4
DECU	1	8	2	9

Note. TA is for “think aloud”.

From Table 46, we can see that there were four to nine tests which had state events. Some tests had only one UPE, and some had more UPEs. Most problems were from the users’ experience, not the evaluators’ thought. Examining the state events for all tests together, we found that the time spent on “talk rather than TA” was significantly longer than the time on thinking aloud and silence (X2=38.093, p<0.001). There was no significant difference between the time spent on thinking aloud and silence (U=0.663, p=0.686). Table 45 also shows that Danish users tended to spend more time on thinking aloud than did Chinese users, whereas Chinese users tended to spend more time on silence than did their Danish counterparts.

Table 47: Users’ point-event behaviors in the UPE period

Point-event behaviors	Result	DEDU	CEDU	CECU	DECU
Negative comments	Total number	2	2	3	16
	Total Rate per minute	1.32	1.71	2.41	6.64
Positive comments	Total number	1	2	1	2
	Total Rate per minute	0.58	1.2	0.52	0.43
Culture related comments	Total number	0	0	0	2
	Total Rate per minute	0	0	0	1.02
Suggestions	Total number	3	0	5	4
	Total Rate per minute	2.4	0	2.13	1.72
Ask questions	Total number	0	1	0	3
	Total Rate per minute	0	0.76	0	3.31

Short reply by the user	Total number	22	10	9	20
	Total Rate per minute	24.92	12.39	4.56	24.09

Table 47 shows the users' behaviors in the UPE period. Comparing Table 44 the users' behaviors in the UPU period, we can see that there were more negative comments than "short reply" (U<0.001, p=0.021) in the UPU period, whereas there were more "short reply" than negative comments (U=1.000, p=0.043) in the UPE period. This implies that when the problems were from the evaluators' thought, the evaluators tended to use lots of probes which just needed the users' short reply of agreement or disagreement. Hence, the users had the behaviors of short reply more than others in the UPE period.

Summary of Theme 4c

The findings in theme 4c are:

1) Users spent significantly longer time on "talking rather than thinking aloud" than thinking aloud and silence in both UPU and UPE periods. Danish users spent significantly more time on thinking aloud than did Chinese users, and Chinese users spent significantly more time on silence than did Danish users in the UPU period. The findings were the same as those in the testing period.
2) In the UPU period, users' negative comments and suggestions per minute were significantly more than those in the testing period. The positive comments per minute were significantly less than those in the testing period.
3) Users gave more negative comments than short replies in the UPU period, whereas users gave more "short reply" than negative comments in the UPE period, which implies that for the UPUs, the users' main role was sending problem information, whereas for the UPEs, the users' main role was receiving problem information.

4.4.5.4 Theme 4d

Theme 4d: Evaluators' communication patterns for the UPU and UPE are different in different cultural settings.

Theme 4d is developed to examine the evaluators' communication patterns in the UPU and UPE periods. This section also includes two parts: communication patterns for the UPUs and communication patterns for the UPEs.

Analysis of the evaluators' point events for the UPUs

Table 48 shows the evaluators' point event behaviors in the UPU period.

Table 48: Evaluators' point-event behaviors in the UPU period

Point event behavior	Results	DEDU	CEDU	CECU	DECU
Affirmative	Total number	202	309	191	211
	Total Rate per minute	59.9	75.32	54.63	57.76
Directed probing	Total number	6	5	9	8
	Total Rate per minute	1.25	1.01	1.89	1.5
Digging deeper probing	Total number	88	55	67	77
	Total Rate per minute	26.34	15.78	16.61	24.1
Using question as the reminder to think aloud	Total number	6	2	4	11
	Total Rate per minute	1.55	0.3	0.74	2.04
Probing: other (general questions)	Total number	13	18	21	12
	Total Rate per minute	3.7	4.74	4.45	3.07
Help behavior	Total number	3	4	10	8
	Total Rate per minute	0.98	0.8	3.45	1.52
No clear answer or help, but encourage	Total number	2	5	1	3
	Total Rate per minute	0.73	1.04	0.2	0.95

Note. TA is for "think aloud".

In Table 48, there were four probing behaviors: directed probing, digging deeper probing, probing as a reminder to think aloud, and probing general questions. The results indicated that the digging deeper probes were significantly more than the other three probes (X_2=12.738, p=0.005) (Kruskal-Wallis Test).

For the affirmative responses, compared to Table 35, it seems that the total rates per minute in the UPU period were more than those in the testing period in all four cultural settings. However, Mann-Whitney Test shows that the difference was not significant (U=4.000, p=0.248). When there were problems, they may have given more affirmative responses but not significantly more than those in the other occasions.

When the users met problems, Chinese evaluators were more willing to provide help than were Danish evaluators, especially when Chinese evaluators were with Chinese users, see Table 49. For Chinese evaluators, in order not to cover the problems met by Danish users, they were not as helpful as when they were with local users.

Table 49: The rates of the evaluators' helping behavior and non-helping behavior in the UPU period

Helping behaviors	Results	DEDU	CEDU	CECU	DECU
Help behavior (A)	Total Rate per minute	0.98	0.8	3.45	1.52
No clear answer or help, but encourage (B)	Total Rate per minute	0.73	1.04	0.2	0.95
Total (A+B)	Total Rate per minute	1.71	1.84	3.65	2.47
Rate of help behavior	A/ (A+B)	0.57	0.43	0.95	0.62
Rate of not willing to help	B/ (A+B)	0.43	0.57	0.05	0.38

Analysis of the evaluators' point events for the UPEs

The evaluators' communication patterns for the UPEs are analyzed in this section. Table 50 shows the evaluators' behaviors in the UPE period.

Table 50: Evaluators' point-event behaviors in the UPE period

Point event behaviors	Results	DEDU	CEDU	CECU	DECU
Affirmative	Total number	24	2	14	58
	Total Rate per minute	17.33	1.71	12.18	31.17
Directed probing	Total number	12	5	6	12
	Total Rate per minute	16.07	8.35	3.66	15.02
Digging deeper or ask for clarifying	Total number	15	3	9	17
	Total Rate per minute	11.16	1.81	3.39	10.16
Using question as the reminder to think aloud	Total number	0	0	0	1
	Total Rate per minute	0	0	0	0.21
Probing: other (general questions)	Total number	1	1	1	0
	Total Rate per minute	0.39	0.6	0.36	0
Help behavior	Total number	0	2	1	2
	Total Rate per minute	0	1.87	1.53	1.69

Note. TA is for "think aloud".

From Table 50, we can see that there were considerable directed probing behaviors. The total rate per minute was much more than that in the UPU period, see Table 48 ($U<0.001$, $p=0.021$). On the other hand, the total rate per minute of digging deeper probing was significantly less than that in the UPU period ($U<0.001$, $p=0.021$). Thus, comparing to the problems experienced by the users first, when they were considered as problems by the evaluators first, there were more directed probes and less digging deeper probes.

Summary of Theme 4d

The findings in theme 4d are:

1) In the UPU period, the digging deeper probes were significantly more than other probes.
2) Chinese evaluators were more willing to provide help than were Danish evaluators, especially when Chinese evaluators were with Chinese users.
3) Evaluators gave more digging deeper probes per minute in the UPU period than in the UPE period, whereas they gave more directed probes per minute in the UPE period than in the UPU period.

4.4.6 Summary of the Communication Analysis

In this section, the evaluators' and users' communication patterns have been analyzed. In the testing period, the time spent on "talking rather than thinking aloud" was significantly longer than the time on thinking aloud and silence. For all users, they spent significantly longer time on talking than thinking aloud or silence. Danish users spent significantly longer time on thinking aloud than on silence, whereas Chinese users spent significantly longer time on silence than on thinking aloud. Danish users spent significantly longer time on thinking aloud than did Chinese users, which implies that Danish users were better at thinking aloud than were Chinese users. In the tests, the Danish users gave significantly more negative comments than Chinese users per minute. Both Danish and Chinese users gave more culture related comments to foreign evaluators than to local evaluators, and the trend was more significant for Chinese users (p_{Duser}=0.054, p_{Cuser}=0.032). Foreign evaluators tended to give more affirmative responses per minute to users than did local evaluators, especially when with Chinese users. All evaluators had considerable digging deeper and directed probing behaviors. The Danish evaluators had significantly more digging deeper probes than did Chinese evaluators. Evaluators tended to use more reminders to make the users talk when with Chinese users than when with Danish users. The reminders were usually in the form of questions instead of the classic reminder "keep talking." Chinese evaluators were more willing to provide help than were Danish evaluators.

Furthermore, the evaluators normally find usability problems based on the users' experience, such as the users' comments, suggestions or task performing behaviors (UPU). However, the evaluators' own thoughts of problems also play a role in finding the problems (UPE). There was no difference of the time spent on the UPU and UPE. For the UPUs, the time spent on minor problems was less than the time on important and critical problems. This trend was especially clear for the tests done by foreign evaluators. The analysis of UPU and UPE show that the state events had a similar trend as those in the tests. Users' negative comments and suggestions and evaluators' digging deeper probes played an important role in finding the UPUs. On the other hand, the evaluators' directed probing and the users' short reply contributed a lot to finding the UPEs.

4.5 Interviews

After the tests, the users and evaluators were interviewed by the researcher (Appendix 6 and Appendix 7). In section 4.5, the findings from the interviews are presented.

4.5.1 Interview of the Users

The users were interviewed by the researcher about speaking English to do the test, and their feelings of doing the test with foreign evaluators.

Nine Danish users said that for them speaking English in the tests was the same as speaking Danish. 23 Danish users and all 32 Chinese users said that doing the tests in English was similar to doing it in their local languages. The major problem for speaking English to do the tests was that it was difficult for the participants who were not familiar with the English names of the MS Word functions to verbalize the functions in English. Another problem was that they could not use too many fancy words to express their feelings in English. However, they all thought they had expressed clearly about their opinions and feelings of the application. The users were also asked whether they would think aloud more and keep silent less if speaking their local languages. All users said that speaking English did not influence the way they were talking. For example, some Chinese users always talked before and after doing the tasks, not during the task performing. In the interview, they said they preferred to do the task without saying anything, regardless of whether the tests were in English or Chinese. In this study, therefore, the influence originating from the English speaking may be limited.

In the interviews, the users were also asked whether they would behave the same to both local and foreign evaluators. The question was:

- If using a foreign evaluator (or since just now the evaluator was a foreign person), which of the following statements fits your situation?

1) I will explain more to the foreign evaluator about what I am doing, since I think he/she may not understand. I hope he/ she can understand why I am choosing this not that.
2) I just do the task and speak out as required. I do not talk more to the foreign evaluator, since he/she cannot understand.
3) Neither. I behave the same to local and foreign evaluators.

Table 51 shows the answers by the users in different cultural settings.

Table 51: The number of the relevant answers in different cultural settings

Answer	DEDU	CEDU	CECU	DECU
Talk more to the foreign evaluators (1)	5	1	10	10
Talk less to the foreign evaluators (2)	0	1	0	0
Behave the same to local and foreign evaluators (3)	11	14	6	6

From Table 51 we can see that most Danish users thought they behaved the same to local and foreign evaluators, especially for Danish users with foreign (Chinese) evaluators. 6 Danish users thought they might talk more to foreign evaluators. However, for Danish users who were really with foreign evaluators, actually only one user thought she talked more. Most Danish users said: "If the foreign evaluator did not understand, he/she would ask, or else, I would expect he/she knew what I said and talk the same thing to both local and foreign evaluators."

In all cultural settings, only one Danish user chose the second one: "I do not talk more to the foreign evaluators." In the interview with the researcher, he said, "I may have some small talk with the local evaluator, such as the traditional Danish marriage that my grandparents had. Since the foreign evaluators did not understand that and it was not easy to explain, I did not say too much about it just now."

Chinese users preferred to talk more to foreign evaluators than to local evaluators, in order to make sure the foreign evaluators could understand.

Section 4.5.1 has presented the findings of the interviews with the users. The interviews of the evaluators are discussed next.

4.5.2 Interview of the Evaluators

In order to gain further insights into the evaluators' problem finding and severity rating behaviors, and the culture's impact on the usability testing, the evaluators were interviewed by the researcher after the tests.

Criteria of deciding what are usability problems

Table 52 shows the criteria of deciding what are usability problems mentioned by Danish and Chinese evaluators.

Table 52: The criteria of deciding usability problems

Item	Criteria description	N. Danish Es	N. Chinese Es
1	From the users' task performing: such as breaking the users' task flow, trying several times to complete the task, not the efficient way to complete the task, taking longer time to finish the task	8	8
2	From the users' comments and suggestions: such as negative comments and suggestions, feelings, preference	6	5
3	Not the same as the users' expectations: such as the functions or outcomes were not the one the users expected	3	3
4	Based on the requirement of the clients, such as what the clients' investment goal was	3	2
5	Purpose of the application, such as whether this issue was important for this application	3	2
6	Based on the evaluators' own experience, expectations.	3	1
7	The frequency of the problem: such as the number of the users having this problem, or the number of issues mentioned by one user. For example, if one user mentioned an issue, the evaluator may not take it as a problem, but if more users mentioned the same issue, it would be taken as a problem.	1	1

Note. N refers to the number of evaluators who mentioned the criterion. E refers to the evaluator

From Table 52, we can see that the users' task performing behavior was important in order for evaluators to detect the problems. The users' task performing could show whether there were any flaws of the application. When users made errors and the task flow broke down, or when the users were unable to complete the tasks efficiently and effectively, usability problems could exist. Even though there was no problem in the task performing, if the user gave some comments or suggestions, the evaluator could also take them as problems.

Moreover, whether it was a problem or not depends on the clients' requirement and the purpose of the application. For example, some evaluators said they might focus only on the utility of the application, not the users' satisfaction of the provided images, texts and backgrounds if the researcher did not introduce the purpose of the application carefully.

Users' and evaluators' expectations and experiences also played an important role in finding the problems. In the interview, one evaluator said, "I anticipated some problems before the tests. If the user did not have this problem, I would ask him/her." Another evaluator said, "Since all the previous users had a specific problem, if this one did not have, I would probe questions to see why he did not have this problem." Table 52 shows that more Danish evaluators than Chinese evaluators took the expectation as one of the criteria of detecting problems, and thus Danish evaluators may be more "active" in the tests than were Chinese evaluators.

Two evaluators also brought forward the criterion of "the frequency of the problem." This criterion may not play a very important role in finding the problems, but it does play an important role in deciding the severity of the problems.

Criteria of rating the severity of the usability problems

The criteria of deciding the severity of the problems are actually related to the criteria of deciding on the problems, only with different severities. For example, for the users' task performing in Table 52, the criteria of critical problems are normally described as "the task cannot be completed" or "it takes too long time to figure out the problem." The criteria of the important problems are normally described as "the task is not easy to do, but the user can find the solution after a while," or "the user needs to try several different options before solving the task." The criterion of minor problems could be "the task is not intuitive to do, but the user can figure it out quickly."

The criteria in Table 52 played a role in deciding the severity of the problems. One more criterion was put forward by the evaluators in rating the problem which was "the influence in the redesign." Table 53 shows the number of the criteria mentioned by the evaluators in deciding on the problem severities.

Table 53: The number of the criteria in deciding the severity of the usability problems

Item	Criteria description	N. Danish Es	N. Chinese Es
1	From the users' task performing	8	8
2	From the users' comments and suggestions	3	4
3	Not the same as the users' expectations	1	0
4	Based on the requirement of the clients	4	2
5	Purpose of the application	5	6
6	Based on the evaluators' own experience, expectations.	4	3
7	The frequency of the problem	5	5
8	The influence in the redesign: such as the cost/ effort to fix the problem for the developers	3	1

Note. N refers to the number of evaluators who mentioned the criterion. E refers to the evaluator.

Comparing Table 53 and 52, we can see that the problem detecting and problem severity rating share similar criteria. However, the importance of the criteria is not exactly the same. For example, item 2 and 3 were mentioned by more evaluators in Table 52 than in Table 53, whereas item 4 to item 7 were used more in deciding on the severity of the problems. Item 8 might not be considered in finding the problems, but it was considered by some evaluators in rating the severity of the problems.

From the interview, we found that when deciding on the severity of the problems, the evaluators often considered several criteria together. For example, if an issue related to the main purpose of the application, even though it could be figured out quickly, it was still a critical or important problem.

From the above findings of the evaluators' criteria of deciding the problems and the problem severities, we can see that culture is not put forward as an independent item. However, culture cannot be ignored, since many criteria involve the evaluators' and users' cultural backgrounds, such as the evaluators' experience in finding and rating the problems (item 6). Item 1 to item 3 may also have been influenced by the evaluators' and users' cultural backgrounds. In the interview, the evaluators were thus asked about the culture's impact on the thinking aloud usability testing.

Culture's impact on the thinking aloud usability testing

In this study, there were 12 evaluators who had had the experience of doing the usability testing abroad ("abroad" here refers to East Asian countries for Danish evaluators and Western countries for Chinese evaluators). Some had gained the experience in this research and some gained it from the experience in their years of being a usability specialist. All thought there were no big differences between the tests in their own culture and in another culture. In the interview, they said usability specialists were educated in the similar usability traditions, and they had been trained in a similar way of doing usability testing. Thus, regardless of where the usability testing was, they would follow similar goals, instructions and guidelines to do the tests. However, they also agreed that there were some differences when doing tests in different cultures and some issues should be considered:

- The first issue was language. The evaluators thought that if they could speak the same language as the target users, there might be no problem for them to do the tests and to find the usability problems.
- Second, when doing tests with local users, the evaluators' own judgment of usability problems and problem severities played a more important role than when doing tests with foreign users. The previous sections showed that some evaluators relied on their own experience or expectations to decide on the problems and the severity of the problems. However, two Danish evaluators who did the tests in China said that their expectations would play a more important role in the tests in Denmark than in China. When with Danish users, even though very few people had this problem, if the evaluators thought it was a problem, they would write it down. But in an unfamiliar culture, even though they thought it was an important problem, if no user or very few users mentioned it, they would not take it as a problem or at least not an important problem. Therefore, the evaluators seemed more "objective" when doing tests in a foreign country.
- Third, because of sharing the same cultural backgrounds, it was easier to interact with local users than with foreign users. A Chinese evaluator commented, "Sometimes I could see that the user wanted to say something but did not say it. If it was a Chinese user, even though he/she did not say it, I would understand the user's meaning and say the sentence for him/her to see whether it was his meaning. However, if it was a Danish user, it was hard for me to say it for him/her, since I did not know what he/she wanted to say."

- Fourth, some evaluators mentioned that Danish users were more talkative in tests than were Chinese users, and thus the way to communicate with the users was not exactly the same. When doing tests with Chinese users, the evaluators had to probe more questions in order to make the users talk. The evaluators' communication skill seemed more important when with Chinese users than when with Danish users.
- Fifth, the evaluators who had a lot of experience in doing tests in different cultures said that some cultural issues, such as attitudes toward males and females, some specific gestures, etc., should be considered when doing tests in some specific cultures, such as Indian or Muslim cultures.

From the interviews, we can conclude that apart from the language issue, culture does have some influence on both the usability problems and communications from the usability specialists' perspective. It is thus worthwhile doing this research to explore out the "concealed" cultural differences which may impact usability testing. When doing tests in another culture, the evaluators' strategy to decide on the problems or the problem severities was not exactly the same as the one in their own culture. They may have relied less on their own judgment. Evaluators might find it easier to communicate with local users than with foreign users, even though they had to speak English in both situations. Because of the users' different preferences in talking while doing the tasks, the evaluators had to communicate differently with the Danish and Chinese users. In addition to the experience from the tests in this study, the evaluators who had experience in doing tests in other cultures also mentioned other issues that should be considered before the tests.

4.5.3 Summary of the Interviews

In section 4.5, the users' and evaluators' interviews have been discussed. All users thought that they had expressed their thoughts and feelings clearly and thoroughly in English. The users' way of talking in English was the same as in their local languages from the users' perspective. Most Danish users thought they would behave the same to local and foreign evaluators, whereas most Chinese users thought they would try to explain more to foreign evaluators than to local evaluators. The evaluators' criteria of deciding on the usability problems and the problem severities were also summarized. Apart from learning problems from the users, the evaluators' own experience and the clients' requirement also played important roles in finding and rating the problems. Even though for usability specialists, the usability testing might be similar all over the world, they still agreed that culture did have some influence in the testing.

4.6 Chapter Summary

Chapter 4 has presented the results in this study. The demographic analysis showed that Danish and Chinese participants represented the target people to some extent and that they were comparable in different cultural settings.

From the problem finding and problem severity rating analysis, we found that, compared to the problem finding behavior, the problem severity rating behavior was more sensitive to the evaluators' and users' cultural backgrounds. Local and foreign evaluators' problem detection rates were similar, and the users tended to agree on most problems found by both local and foreign evaluators. Chinese evaluators preferred to rate the problem severity as "important" more than "minor" or "critical," whereas Danish evaluators did not have such tendency. However, when doing tests with foreign users, the evaluators' tendencies on rating the problems were changed. Both Danish and Chinese evaluators tended to rate less "critical" than "important" and "minor." Local evaluators tended to give similar ranks as the target users. However, when doing tests with Danish users, Chinese evaluators gave similar severity ranks as did Danish users.

The communication analysis showed that there were some differences for the users' and evaluators' communication patterns in different cultural settings, the findings of which have been summarized in section 4.4.6. From the analyses of the identified usability problems and the communication patterns, we concluded that culture's impact on the process of the usability testing (communication) was greater than the results of the usability testing (usability problems). Finally, the interviews with users and evaluators after the tests were discussed (see the summary in section 4.5.3).

The results are discussed in the next chapter.

5 Discussion

This chapter discusses the findings in this study and includes five sections: 1) Participants in this research; 2) Discussion of the first sub-research question: examining hypotheses 1 to 4 regarding the results of the thinking aloud usability testing-identified usability problems; 3) Discussion of the second sub-research question: investigating the themes on the process of the thinking aloud usability testing- communications; 4) Evaluators' views of thinking aloud usability testing; 5) Summary.

5.1 Participants in this research

Since the independent variable is the evaluators' and users' cultural backgrounds, selecting the right participants is very important in order for this research to obtain reliable and valid results. This section discusses the demographic information of the evaluators and users, as well as the interview with the users in order to see whether the participants are representative and comparable in different cultural settings.

Usability specialists who had more than 1 year experience in the usability engineering field and had done usability tests before were recruited as evaluators in this study. The evaluators' ages ranged from 25 to 40 years old. The years of education were similar for the evaluators in the four cultural settings. Moreover, they had similar experience of using MS Word and were able to speak English fluently. Thus, the evaluators in this study were all similarly skilled usability practitioners and therefore were assumed to be comparable in different cultural settings. Regarding the demographics of the users in Denmark/ China, their ages, years of education, English skill and familiarity with MS Word were also comparable. Hence, in this research, if the Danish and Chinese evaluators found different problems when with the same target users, it could be assumed that it was because of the differences in the evaluators' cultural backgrounds, not other factors (such as spoken English skills, familiarity with MS Word, experience of being usability professionals).

This researcher attempted to find an equal number and gender of evaluators and users, that is, half male and half female. However, because of the difficulty in recruiting participants with good spoken English, as well as willingness to attend this study, the gender was not always the same in different cultural settings. According to the study conducted by Clemmensen et al. (2007), gender is not a big issue for usability tests in Denmark and China. From the interviews with the evaluators in this research, it is clear that none regarded gender as an issue in the tests; however, some did mention that they might consider gender to be an issue if they were doing tests in Indian or Muslem cultures.

All participants in this study were good at English, and at the interviews, all users thought they could express their thoughts clearly and that their communication had not been affected by speaking English. We conclude that the identified problems and the communication patterns should not have been impacted by the English language and that the findings from this study, to some extent, should be able to be generalized to situations with local languages.

All local evaluators and users were born and raised in the target country (Denmark or China); all foreign evaluators were invited from China/ Denmark for this study, and their travel fees were paid by the project. They were assumed to be typical Danish or Chinese, as described by culture theories (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2003; Nisbett, 2004; Nisbett & Masuda, 2003; Nisbett & Norenzayan, 2002a). Of course, as described in section 3.1.2.1 and section 3.2.1, not all Danish people have an analytic cognitive style and low contextual communication, and not all Chinese people have holistic cognitive style and high contextual communication. The focus of this research is the cultural variation approaches to usability testing. Nisbett's and Hall's studies have shown Danish and Chinese people have these orientations, and in this research all participants were recruited from the target countries; we could therefore assume the participants to have these tendencies. This research investigates Danish and Chinese evaluators' and users' behaviors at the group level. (The limitations of this research are discussed in section 6.3.)

Two questions related to the cognitive style of categorization were used as detectors to check the participants' cultural orientations. The answers showed that Chinese participants had the tendency of grouping objects according to the relation and family resemblance, whereas Danish participants had the tendency of grouping objects according to the features of the objects and the rules. Apart from the two questions, in the interviews (section 4.5.1), most Danish users reported that they would behave the same with local and foreign evaluators, whereas most Chinese users reported explaining more to foreign evaluators in order to help them understand. This finding shows the difference between Danish and Chinese participants, and complies with the task-focus and socio-emotional orientations in section 2.2.2.2. Thus, the participants in this study could represent the target people described by the culture theories at the group level.

In the next sections, the results of the hypotheses and themes examined by this dissertation will be discussed. Since the influences of the cultural differences—cognition and communication on usability testing—were discussed in detail in section 2.2 of the theoretical background chapter, these influences will not be repeated here. The discussion is limited to the findings in this dissertation.

Moreover, in this discussion chapter, the words "Western" and "East Asian" will be used in many places instead of "Danish"/ "Chinese" in order to have a better understanding of the culture theories in this research (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Norenzayan, 2002a). As discussed in section 3.2.1, Denmark is used to represent the Western culture, and China is used to represent the East Asian culture. The findings and discussions in this research should be generalizable to a broader area.

5.2 The First Sub-Research Question---Hypothesis 1 to Hypothesis 4

The first sub-research question is regarding usability problem finding and problem severity rating behaviors in different cultural settings. There was no significant difference for the problem finding behaviors in different cultural settings, whereas there were some differences for the problem severity rating behaviors. The discussion of the results is organized into two sections: 1) usability problem finding behaviors which have no significant difference, involving Hypothesis 1 and Hypothesis 3, and 2) problem severity rating behaviors which have some differences as theoretically predicted, involving Hypothesis 2 and Hypothesis 4.

5.2.1 Usability Problem Finding Behaviors (Hypotheses are rejected)

5.2.1.1 Hypothesis 1

Hypothesis 1: The usability problems found by local evaluators are different from the usability problems found by foreign evaluators: H1 a (in Denmark): When with Danish users, the usability problems found by Danish evaluators are different from the usability problems found by Chinese evaluators; H1 b (in China): When with Chinese users, the usability problems found by Chinese evaluators are different from the usability problems found by Danish evaluators.

Hypothesis 1 is rejected. There was no significant difference for the Western and East Asian evaluators' identified usability problems when with the same target users. Different cognitive styles and communication orientations did not seem to influence the identified problems in the thinking aloud usability testing. In other words, problem identification did not seem to be sensitive to the evaluators' cultural backgrounds. Possible reasons are that: 1) all evaluators were professionals and had the ability to detect problems; 2) there were 16 tests in each cultural setting, and it could be assumed that most potential problems were found and it was hard to find more problems. In this regard, researchers (Law & Hvanneberg, 2002; Nielsen & Landauer, 1993) have purported that five tests are necessary to capture 80% of the known usability problems of a system. Even though the "magic five" is questioned by some researchers (Law & Hvanneberg, 2004; Woolrych & Cockton), the doubts usually focus on the diverse contexts of the testing, such as different users, different test conditions, etc. In this study, however, the evaluators used the same protocol for the testing and the users were similar, which meant that the variation in the test setting was small, for which reason the 16 evaluators in each cultural setting were able to find most usability problems. Thus, if all other factors were the same, with the only difference being the evaluators' cultural background, as long as there were sufficient tests, the identified problems by the local and foreign evaluators would not be significantly different.

There is a difference in the results of this study and those of the work of Vatrapu and Pérez-Quiñones (2006), who reported that local evaluators found more usability problems than foreign evaluators did when with Indian users in the US. The reasons for this difference may be: 1) they investigated structured interviews, not thinking aloud usability testing. Finding usability problems in thinking aloud usability testing relies not only on users' comments, but also on users' task performing. 2) The locations of the usability tests were different. The tests in their research were conducted with Indian users in the US, whereas the tests in this study were conducted with target users in target countries. The foreign evaluators were invited to the target countries just for the tests; for this reason, they may have taken the tests more seriously and were able to find as many problems as did the local evaluators.

Even though from the quantitative analysis there was no significant difference between the local and foreign evaluators' identified problems, an analysis of the unique problems and shared problems could help our understanding of the cultural impact on finding usability problems in usability testing. The following sections discuss the unique problems and shared problems.

Discussion of the unique usability problems

As discussed earlier, there were sufficient tests, and the evaluators were usability professionals, which could suggest that there was no significant difference with the problems found by Western and East Asian evaluators. However, the finding is not exactly the same for the tests with Western and East Asian users when analyzing the identified problems qualitatively. For the tests with Western users, there were no unique problems, that is, all problems found by Western evaluators were also found by East Asian evaluators. But for the tests with East Asian users, there were 6 unique problems. Three problems were found only by East Asian evaluators, and three found only by Western evaluators.

Given this finding, we cannot say whether the evaluators' cultural backgrounds have an impact on the identified usability problems in the thinking aloud usability testing or not. The users' cultural backgrounds should also be considered. Western users who tend to have low-contextual communication and task-focus orientation may have the tendency of expressing everything in their mind, regardless of being with local or foreign evaluators. In the interviews (section 4.5.1), most Danish users said they would behave the same with local and foreign evaluators. Thus, through a sufficient number of tests, the foreign usability specialists were able to find as many usability problems as the local usability specialists found.

On the other hand, East Asian users, who tend to have high-contextual communication and socio-emotional orientation, may not behave the same with local and foreign evaluators, which could result in some differences for the local or foreign evaluators' identified problems, even though the difference is not statistically significant. In the interviews (section 4.5.1), most Chinese users said that they preferred to explain cultural issues to foreign evaluators since foreign evaluators might not know about the culture, and telling them about it would help them understand. In contrast, they seldom spoke about cultural issues to local evaluators since they supposedly knew their own cultural issues, and it was therefore not necessary for users to repeat it. Moreover, Chinese users tend to use high-contextual communication and therefore may not express everything if they felt that the evaluator already knew the information. On the other hand, with foreign evaluators, the users would give contextual information. The observed empirical evidence was similar to the theoretical prediction. Two of the three unique problems which were found only by Danish evaluators when with Chinese users were actually similar to Chinese cultural issues. One was "some images are not used for Chinese wedding invitation" and the other was “problem of making the text vertical.” From this finding, it seems that foreign evaluators elicited the Chinese users’ comments or behaviours related to culture.

A further implication is that the unique problems in the Chinese testing settings influencing the evaluator’s own opinions and feelings of the application may have contributed to the result of the usability problems (Jacobsen, 1999); this may lead to a better understanding of the evaluator effect in the usability testing (Hertzum & Jacobsen, 2001) (section 2.3.1.1). From the unique problem description in section 4.2.1, we can see that the three unique problems which had been found by Chinese evaluators were found only by one Danish evaluator in one test when with Danish users, implying that these were uncommon problems for Danish people, and so it was hard for Danish evaluators to detect them. The finding indicates that although usability testing is a more objective UEM compared to other methods, and it is often used to verify other UEMs (Hartson et al., 2001; Hvannberg et al., 2007; Yeo, 2001b), it is not as objective as researchers have thought, and the evaluators’ own mind also plays an important role in finding usability problems.

Discussion of the shared usability problems

From the analysis of the shared usability problems in each cultural setting, we found that compared to local evaluators, foreign evaluators tended to have more problems found by all four evaluators, and the question remains of why foreign evaluators shared more problems than local evaluators did. In the study conducted by Nørgaard and Hornbæk (2006), the authors found that evaluators asked questions about nonexistent parts of the system, speculative or hypothetical questions, implying that the evaluator’s own mind may call the user’s attention or guide the user’s direction. Compared to foreign evaluators, local evaluators might have more confidence to rely on their own experiences and feelings in finding problems. From the interviews in section 4.5.2, some evaluators proposed “evaluators’ own experience and expectations” as one of the criteria for deciding on the problems. However, they also said that if they were doing tests in a foreign country, they would rely less on their own experiences since they were not from the target user group. When learning problems from users where users are similar, such as in this study (the users had similar years of education and similar experience with the application and did the same tasks), there may be more possibility that all evaluators would find similar problems after four tests. On the other hand, for local evaluators, since they may rely more on their own minds, the problems shared by all evaluators may be less.

Additional discussion

This study also examined the usability problems found by evaluators within each cultural setting. The result shows that the evaluators' problem detection rates for the Danish-Danish pairs were significantly different. Even though there might be individual differences, the main focus in this study is to see whether there is any systematic difference between the evaluators' problem finding behaviors in different cultural settings, and so the individual difference inside a cultural setting may not influence the comparison of the identified problems in different cultural settings. The individual difference in this study indicates that sometimes the individual difference may be greater than the cultural difference. Culture's role in finding usability problems in the usability testing may be limited.

Summary

The comparison of the usability problems found by the evaluators with different cultural backgrounds indicates that culture's impact on problem identification in thinking aloud usability testing is limited. Foreign usability specialists are thus able to find a similar number and similar types of problems as the local usability specialists do when there are sufficient test sessions, especially for tests with Western users. The discussion of the unique problems and shared problems implicates that the interaction between users and evaluators may play an important role in finding usability problems. It may be easier for foreign evaluators to do tests with Western users, whereas it may be more difficult for foreign evaluators to do tests with East Asian users.

5.2.1.2 Hypothesis 3

Hypothesis 3: The usability problems found by local evaluators are more consistent with the usability problems found by target users, compared to the usability problems found by the foreign evaluators and target users: H3 a (in Denmark): The usability problems found by Danish evaluators are more consistent with the usability problems found by Danish users, compared to the usability problems found by Chinese evaluators and Danish users; H3 b (in China): The usability problems found by Chinese evaluators are more consistent with the usability problems found by Chinese users, compared to the usability problems found by Danish evaluators and Chinese users.

Hypothesis 3 is rejected. The hypothesis supposed that if the evaluators and users tended to have similar cognitive styles and communication orientations, there would be more opportunities to find similar problems, compared to the evaluators and users who had different cognitive styles and communication orientations. However, this hypothesis is not supported by the data in this study. Both Western and East Asian users tended to agree with most problems found by both local and foreign evaluators.

The main reason may be that the research design for investigating this hypothesis was not appropriate. This hypothesis is investigated based on the original problem list (section 3.5.2). As described in section 3.5.2.1, considering the users' ability to find usability problems, the users were asked to mark the problems on the list, and the analysis focused on examining whether the usability problems detected by the evaluators were also regarded as problems for users. In this study, all users tended to mark many more problems than the problems found by evaluators. The reasons may be: 1) the way to find problems was different for evaluators and users. Using the list which included most potential problems may have encouraged users to choose more problems than they really talked about in the test. However, for evaluators, since they had to write down the problem, if the problem was not mentioned in the test, it could not be brought forward by evaluators. 2) The understanding of the usability problems was different. For users, even though it may have been a very small problem, when they saw the description on the list, they may have marked it. However, for the evaluators, if it was not a clear problem, they might not have written it down until it happened again in the following tests. 3) The analysis did not include the problems found by evaluators but not on the list. Because of the users' ability of finding problems, most of the "other" problems found by evaluators were not found by users, and thus the problems for the analysis are limited to the users' list in the test. Since users tended to mark more problems, there would be more opportunity for them to agree with the problems found by the evaluators. The research design for this hypothesis may need to be improved in the future, such as finding an effective way to elicit the users' thoughts of the problems.

Another reason that both Western and East Asian users tended to agree with most problems found by both local and foreign evaluators may be that not all participants in this study had the same cognitive styles and communication orientations as the culture theories described, or the culture theories in this research were not sufficient. In this research, we did not really test the participants' cognitive styles and communication orientations. As discussed in section 3.1.2.1, section 3.2.1 and section 5.1, the participants' cognitive styles and communication orientations were mainly based on existing culture theories (Hall, 1989a; Hall & Hall, 1990; Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Norenzayan, 2002a). However, not all Danish/Chinese people have cognitive styles or communication orientations as described by the theories. When doing the research at the group level, the identified problems may not be sensitive enough to the participants' cultural backgrounds.

Two solutions could be considered to solve this issue: 1) finding participants with cognitive styles and communication orientations as the theories described; 2) finding additional culture theories to enrich Nisbett's and Hall's theories. For the first solution, we need to do more work on selecting participants in the future. For the second solution, a dynamic constructivist approach to culture, proposed by Hong and Mallorie (2004), is recommended. In Nisbett and his colleagues' studies and Hall's work, culture is assumed to be relatively static and monolithic. However, with globalization of the economy and technology and the transnational dissemination of ideas and languages, people are more and more similar, and there are also more and more poly-cultural/ bicultural people. Sometimes people in different cultures may behave differently, but sometimes they may behave the same. Hong and Mallorie (2004) have proposed a dynamic constructivist approach to culture, focusing on the situation-specific activation of cultural schemata. In this approach, culture's influence is not static, but related to the situation. The situation (or social context) would influence the effect of culture on cognition, affect and behavior. The interactions of "culture×domain" or "culture×situation" are stressed. How well established cross-cultural differences may appear or disappear depends on the availability, accessibility, and applicability of culture theories (Hong & Mallorie, 2004). In the situation of the thinking aloud usability testing, evaluators' and users' cultural backgrounds may play a limited role in finding usability problems.

Summary

Users were in agreement with most usability problems found by evaluators. The tendency was similar for both local and distant pairs, which may be due mainly to the measurements of the users' and evaluators' problems. The research design may need to be reconsidered in future work.

Moreover, in order to have more significant results, all participants should have cognitive styles and communication orientations as the culture theories describe. Furthermore, the dynamic constructivist approach to culture could be considered in the future.

5.2.2 Usability Problem Severity Rating Behaviors (Hypotheses are supported or partially supported)

5.2.2.1 Hypothesis 2

Hypothesis 2: Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems: H2 a (in Denmark): When with Danish users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems; H2 b (in China): When with Chinese users, Danish evaluators and Chinese evaluators have a different tendency in rating the severity of the usability problems.

Both H2a and H2b were supported by the data. Table 54 shows the findings in hypothesis 2.

Table 54: Evaluators' preference in rating the usability problems in different cultural settings

Pairs	Description of the evaluators' tendency in rating the UPs
DE with DU	No preference for rating the UPs as minor, important or critical
CE with DU	Prefer to rate the UPs as minor and important, not critical
CE with CU	Prefer to rate the UPs as important more than minor and critical
DE with CU	Prefer to rate the UPs as minor and important, not critical

Note. DE for Danish evaluators, DU for Danish users, CE for Chinese evaluators, CU for Chinese users, UP for usability problems

From table 54, we can see that Western and East Asian evaluators have different tendencies in rating the severity of usability problems when doing tests in both Western and East Asian countries. From the analysis of the severity ranks given by the evaluators, we found that East Asian evaluators gave more middle ranks to problems, and this tendency was even clearer when with East Asian users. However, Western evaluators did not have such tendency. This finding might be supported by Nisbett's theory (Nisbett, 2003; Nisbett et al., 2001). East Asians are expected to seek compromise solutions to problems and would like to choose the "middle way" solutions for conflicts because of their dialectic reasoning. This feature is reflected in rating the problem severity. If there were only three ranks, East Asian evaluators preferred to choose the middle one. In order to avoid this tendency, we could consider using an even number of ranks, instead of an odd number of ranks, such as four severity ranks, when doing tests with East Asian evaluators.

The result also shows the influence of the users' culture on the evaluators' severity rating behaviour. East Asian evaluators had a preference of rating the problems as "important" more than "minor" or "critical." However, when East Asian evaluators were with Western users, this preference was changed. They tended to rate significantly fewer critical problems. The Western evaluators had similar changes in rating problems when with East Asian users. When with Western users, Western evaluators did not have the tendency of giving minor, important and critical ranks, whereas when with East Asian users, the Western evaluators tended to rate less critical problems. This finding implies that when doing tests with users from different cultures, the evaluators' strategy of deciding on the severity of problems could have changed. They were reluctant to rate the problems as critical, which means that the evaluators tended to be relatively mild in rating the problems when doing tests with foreign users.

Summary

Evaluators with different cultural backgrounds have different tendencies in rating the severity of usability problems. East Asian evaluators have the tendency of giving the middle rank to problems, whereas Western evaluators do not have such tendency, so it is especially important to have an even number of ranks for East Asian evaluators than it is for Western evaluators. Moreover, severity rating behaviours are influenced by culture. Both Western and East Asian evaluators are not willing to give the "critical" rank to the problems when doing tests with foreign users.

5.2.2.2 Hypothesis 4

Hypothesis 4: The usability problem severity ranks given by the local evaluators and target users are more consistent than the problem severity ranks given by the foreign evaluator and target users: H4 a (in Denmark): The usability problem severity ranks given by Danish evaluators and Danish users are more consistent than the problem severity ranks given by Chinese evaluator and Danish users; H4 b (in China): The usability problem severity ranks given by Chinese evaluators and Chinese users are more consistent than the problem severity ranks given by the Danish evaluator and Chinese users.

H4a was not supported by the data, whereas H4b was supported. When with Western users, the severity ranks given by evaluators and users were significantly correlated, regardless of whether with Western evaluators or East Asian evaluators. Since both local and distant pairs had the similar trend, H4a was rejected. On the other hand, when with East Asian users, East Asian pairs' severity ranks were marginally significant (p=0.058), whereas the distant pairs' severity ranks were not significantly correlated (p=0.250); thus, the severity ranks given by the East Asian pairs were more consistent than were those of the distant pairs, and H4b was accepted.

Hypothesis 4 was proposed based on the differences of both cognitive styles and communication orientations discussed in the theoretical background chapter. However, the findings suggest that even though Westerners and East Asians may have different cognitive styles and communication orientations, the impact of the evaluators' cultural backgrounds on problem severity rating was not the same in Western and East Asian testing settings.

As introduced in section 2.2.2.2, Kim (2002) found that thinking aloud was best suited to people with an analytic cognitive style, whereas people with a holistic cognitive style found it difficult to verbalize their thoughts. According to Hall's low-contextual communication and high-contextual communication orientations, Western people tend to speak their mind, whereas East Asian people tend to think that when some information is known to people, it is not necessary to express everything. The findings in this study show that it was easier for foreign evaluators to prompt for Western users' thoughts, but more difficult for foreign evaluators to prompt for East Asian users' thoughts. The findings implicate two important issues in rating usability problems: 1) users' thinking aloud data or comments, and 2) similar cognitive styles and communication orientations shared by evaluators and users. If users are good at thinking aloud, then it may be helpful for evaluators in getting the users' feelings of the usability problem severity. If users are not good at thinking aloud, but they share similar cognitive styles and communication orientations with the evaluators, then the evaluators may also be able to get the users' problem severity. If the users are not good at thinking aloud and they also do not share similar cognitive styles and communication orientations with the evaluators, then it may be more difficult for the evaluators to get the users' problem severity.

Summary

When doing tests with Western users, both Western and East Asian evaluators were able to rate the severity of the usability problems similar to the users, whereas when doing tests with East Asian users, the problem severities given by the local pairs were more consistent than with the problem severities given by distant pairs. The finding suggests that in comparing the tests with East Asian users, it may be easier for foreign evaluators to do tests with Western users who have an analytic cognitive style and low-contextual communication orientation.

5.3 The Second Sub-Research Question---Theme 1 to Theme 4

For the second sub-research question regarding the evaluators' and users' communications, themes instead of hypotheses were developed, given the more explorative nature of the study and the difficulty of developing clear hypotheses beforehand (section 2.3.2.2). Four themes were developed for the second-research question. The previous three themes were the communication patterns in usability testing. As discussed in section 2.3.2.2, the purpose of usability testing is to find usability problems (Nielsen, 1993), making it necessary to investigate the communication patterns for the problems. The fourth theme is regarding the evaluators' and users' communications for the problems. In this section, the findings in the four themes are discussed.

5.3.1.1 Theme 1

Theme 1: The time spent on the users' three state events---thinking aloud, talking rather than thinking aloud, and silence, is different in different cultural settings.

The findings in theme 1 were: 1) for all the pairs, the time on "talking rather than thinking aloud" was significantly longer than the time on thinking aloud or silence; 2) for Western users, the time on thinking aloud was significantly longer than the time on silence; 3) for East Asian users, the time on thinking aloud was significantly less than the time on silence; 4) Western users spent significantly longer time on thinking aloud than did East Asian users, whereas East Asian users spent significantly longer time on silence than did Western users.

The analysis indicates that the users seldom followed the classic thinking aloud model described by Ericsson and Simon (1993). The finding implicates that the third level of verbalization, which is regarded as the least reliable in Ericsson and Simon's model, is actually the most important data in the usability testing (section 2.1.3.1). The users prefer to give the information which comes from the long term memory, such as comments, feelings and suggestions.

Furthermore, compared to East Asian users, Western users spent longer time on thinking aloud, whereas less time was spent on silence. The finding may be related to the differences of cognitive styles and communication orientations discussed in section 2.2.2 and section 2.2.3. It also supports the study conducted by Kim (2002), which showed that East Asians were not as good at thinking aloud as were Westerners, and that silence was beneficial for high levels of thinking for East Asians.

Summary

The discussion in theme 1 indicates that in the thinking aloud usability testing, all users seldom followed the classic thinking aloud model, but instead, they communicated with the evaluators. The "thinking aloud" in the usability testing is closer to the description of Boren and Ramey's model (2000). Moreover, Western users were better at thinking aloud than were East Asian users, and East Asian users spent more time on silence than did Western users. This suggests that Western and East Asian users, who tend to hold different cognitive styles and communication orientations, may have different abilities in thinking aloud and have different preferences in silence.

5.3.1.2 Theme 2

Theme 2: Users' verbal behaviors are different in different cultural settings.

The findings in theme 2 were: 1) Western users gave significantly more negative comments than did East Asian users; 2) users gave more culture related comments to foreign evaluators than to local evaluators. The tendency was especially stronger for East Asian users; 3) there was no significant difference for the positive comments and suggestions given by users in different cultural settings.

According to the task- focus and socio-emotional orientation in Nisbett's theory (section 2.2.2.2), Westerners have a task-focus orientation, whereas East Asians have a socio-emotional orientation. East Asian users might not be willing to give many negative comments in order to establish a harmonic relationship with the evaluators. The finding was similar to previous studies conducted by Lee and Lee (2007) and Yeo (1998a, 2001b). Even though in the field study (Shi, 2008a), East Asian users thought that they had given all negative comments in the tests, compared to Western users in this research, they were less negative.

Further, both Western and East Asian users gave more culture related comments to foreign evaluators than they did to local evaluators. The foreign evaluators' selves acted as stimulators who then activated the users' knowledge of the "cultural issues." When doing tests in a target country with local evaluators, it seems weird to talk about "culture." However, when doing tests in a target country with foreign evaluators, the users' culture related knowledge may be activated. The finding differs from the study done by Vatrapu and Pérez-Quiñones (2006), who found that when with local evaluators, users tended to give more culture related comments. This may be due to the fact that their research was conducted in the US with Indian users, where the location was not the target country that the users were from. In the tests, when Indian users met Indian evaluators in the US, local evaluators acted as stimulators in activating the Indian users' culture related knowledge. The location difference may have influenced the findings in the two studies. When doing tests in the target country with local evaluators, the users may not consider the cultural issues. However, when users are with foreign evaluators in the target country, they may think about cultural issues and explain their feelings regarding the application.

Furthermore, although both East Asian users and Western users tended to give more culture related comments to foreign evaluators, the tendency was stronger for East Asian users. This may also be related to the task-focus vs. socio-emotional orientations. As discussed earlier, East Asians have a socio-emotional orientation, which implies that they may behave differently to the local and foreign evaluators, as compared to Western users.

Summary

From the discussion of theme 2, we learnt that users with different cultural backgrounds may have different attitudes towards negative comments. In the usability testing, even though the users understood the purpose of the tests and knew that negative comments would not make anybody uncomfortable, East Asian users were still not as negative as Western users were. This suggests that usability practitioners should be aware of East Asians' socio-emotional orientation, and should identify the problems based on both their comments and their behaviours. Moreover, when doing usability tests in the target country with target users, foreign evaluators may find it easier to elicit users' culture related comments than do local evaluators.

5.3.1.3 Theme 3

Theme 3: Evaluators' verbal behaviors are different in different cultural settings.

The findings in theme 3 were: 1) when with East Asian users, Western evaluators gave more affirmative responses than did East Asian evaluators, and it was marginally significant; 2) Western evaluators gave significantly more digging deeper probes than did East Asian evaluators. The tendency was especially significant when with Western users; 3) Evaluators used significantly more non-classic reminders than classic reminders to keep the users talking; evaluators gave more reminders to East Asian users than to Western users; when with Western users, Western evaluators gave more reminders than did East Asian evaluators; 4) East Asian evaluators seemed more willing to provide help than were Western evaluators.

Affirmative responses indicated that evaluators were listening actively rather than passively. According to Nisbett et al. (2001), Westerners tend to hold a more active attitude in dealing with their environment, compared to East Asians (section 2.2.2.2). The finding in this study supports this theory. The results of this study show that all evaluators were listening actively; however, Western evaluators were more active than were East Asian evaluators.

Moreover, compared to East Asian evaluators, Western evaluators were more willing to probe digging deeper questions in order to ask the users to clarify their meaning. This finding may be related to the Western users' low-contextual communication orientation. People with low-contextual communication need to be given all the detailed information in order to understand the partner's meaning. Western evaluators, compared to East Asian evaluators, may have less tolerance if given unclear or insufficient information, and thus they preferred to give more digging deeper probes to figure them out.

The classic reminders described by Ericsson and Simon (1993), such as "please keep talking" or "please remember to think aloud," were seldom used in the usability testing settings. Non-classic reminders, such as "how do you think of it?" or "and now…?" which were mentioned by Boren and Ramey (2000) were used often in the tests. Moreover, all evaluators gave more reminders to East Asian users, indicating that East Asian users might easily forget to talk without the evaluators' reminders. Furthermore, regarding the tests with Western users, Western evaluators tended to give more reminders than did East Asian evaluators, which implies that East Asian evaluators might have more tolerance for silence than do Western evaluators.

Compared to Western evaluators, East Asian evaluators were more willing to help users when they had problems. According to Nisbett's theory, East Asians tend to have the socio-emotional orientation. In order to make users feel comfortable and not frustrated, East Asian evaluators may be willing to provide help.

Summary

The discussion of theme 3 has shown that Western evaluators were more active in probing questions and giving reminders than were East Asian evaluators, which suggests that they may have less tolerance in unclear or insufficient information and silence than do East Asian evaluators. On the other hand, East Asian evaluators were more active than Western evaluators were in providing help, which may originate from their socio-emotional orientation. Moreover, more reminders were given to East Asian users than to their Western counterparts. In theme 1, East Asian users spent longer time on silence than did Western users; thus, the finding here shows an interaction between the users' state events and the evaluators' behaviours.

5.3.1.4 Theme 4

Theme 4: The way to communicate usability problems is different in different cultural settings.

Theme 4 is related to the communication patterns for the usability problems. In this study, the problems were divided into UPU and UPE. UPU and UPE involve different ways of transmitting problem information. The problem information in UPU is transmitted from the user to the evaluator, whereas the problem information in UPE is transmitted from the evaluator to the user. From the lag sequential analysis in section 4.4.5, we can see that UPU and UPE have different behaviors in starting the problems. UPU normally starts with the users' negative comments or suggestions, whereas UPE normally starts with the evaluators' directed probes. Since the UPU and UPE may involve different communication patterns, they were analyzed separately.

There are four sub-themes in theme 4. In the following sub-sections, the four sub-themes are discussed individually.

Theme 4a: The number and duration of the UPU and UPE may be different in different cultural settings.

The findings in theme 4a were: 1) there were significantly more UPUs than UPEs in all cultural settings; 2) the rates of Western evaluators' UPEs (the number of UPEs found by the Western evaluators/ all the problems found by the Western evaluators) were two or three times larger than the rates of the East Asian evaluators' UPEs; 3) there was no significant difference for the durations of UPUs and UPEs in all the cultural settings.

The finding of more UPUs than UPEs indicates that in the usability testing, the problems were mainly learnt from the users, and the evaluator's main role was being the learner. Both Ericsson and Simon's thinking aloud model (1993) and Boren and Ramey's thinking aloud model (2000) emphasize the users' role in detecting problems. According to Ericsson and Simon's model (1993), the evaluators are passive listeners, and according to Boren and Ramey's model (2000), the evaluator's main role is being the learner, and the user is the expert who is assumed to provide the information about the problems. Based on the two thinking aloud models, the problems should be UPUs not UPEs. However, in the usability testing, the evaluator was not always the receiver of the problem information. The existence of UPEs indicates that sometimes the evaluator may send problem information, such as using probes to direct the user's attention to specific issues that the user did not notice (Shi, 2008a; Tamler, 2003).

The rates of Western evaluators' UPEs were larger than that of East Asian evaluators. According to the theory of perception of control (Ji et al., 2000; Nisbett et al., 2001) (section 2.2.2.2), Western evaluators may be more active than are East Asian evaluators in starting a conversation to check the potential problems, whereas East Asian evaluators may depend more on observing the users' task performing behaviours and listening to users' comments.

The durations of the UPUs and UPEs had no significant difference in all cultural settings. Participants with different cultural backgrounds may tend to spend similar time on discussing the UPUs and UPEs.

Theme 4b: The time spent on discussing UPU/ UPE with different severity is different in the four cultural settings.

The findings in theme 4b were: 1) for the UPEs and UPUs found by local evaluators, there was no relation between the time on the problems and the severity of the problems. 2) For the distant pairs' UPUs, the foreign evaluators spent significantly more time on critical problems than on minor problems, and the tendency was especially significant for the Western evaluator-East Asian user pairs.

Problem handling time is regarded as a way to prioritize problems (Hassenzahl, 2000) (section 2.3.1.1 and section 2.3.2.2). The analysis showed that for the UPUs, foreign evaluators tended to spend more time on critical problems than on important problems, and spend more time on important problems than on minor problems. From the interviews after the tests (section 4.5.2 and section 5.4), we found that the evaluators' own experience was regarded as one of the criteria to decide on the usability problems and the severity of the problems. However, the evaluators also reported that they relied less on their own experience, and more on the users' task performing behavior when doing tests in another country. The time spent on problems could be regarded as being one of the important measures in the users' task performing behavior. The UPEs and also the local evaluators' UPUs involved the evaluators' own judgment of the severity, and so there was no consistent tendency for the time on the problems with different severities. On the other hand, foreign evaluators who were not from the target user group and who were learning the problems from the users may take the more objective measure---time, as one of the factors to decide on the severity of the problems. They may rate the severity of the problems based on the time spent on discussing or dealing with the problems. The result implies that when evaluators are not familiar with problems, they may tend to rely more on the objective measures (Hassenzahl, 2000) to decide on the severity, whereas when the evaluators have had some understanding of the problems, they may rely less on the objective measure.

Furthermore, for UPUs found by foreign evaluators, Western evaluators (when with East Asian users) relied more on objective measures than did East Asian evaluators (when with Western users). In the Western evaluator-East Asian user setting, East Asian users tended not to be good at thinking aloud, and thus in order to decide on the severity of the problems, Western evaluators tended to rely more on the objective measure-time.

Theme 4c: Users' communication patterns for the UPU and UPE are different in different cultural settings.

The findings in theme 4c were: 1) Users spent significantly longer time on "talking rather than thinking aloud" than thinking aloud and silence in both UPU and UPE periods. Western users spent significantly more time on thinking aloud than did East Asian users, and East Asian users spent significantly more time on silence than did Western users in the UPU period; 2) In the UPU period, users' negative comments and suggestions per minute were significantly more than those in the testing period; the positive comments per minute were significantly less than those in the testing period; 3) Users gave more negative comments than short reply in the UPU period, whereas users gave more short replies than negative comments in the UPE period; 4) For all users' point event behaviours, there was no significant difference between the local and distant pairs, and there was also no significant difference between the Western and East Asian pairs.

The tendencies for the users' three state events were the same as those in the testing period. In the usability problem period, Western users spent more time on thinking aloud than East Asian users, and East Asian users spent more time on silence than Western users.

Regarding the users' point event behaviours, in the UPU period, negative comments and suggestions were important signs for indicating problems. A negative comment showed the negative aspect of the application or the trouble of the task performing. The suggestion was often given by users as the improvement of the application or the users' wish of using the application in an easier or more convenient way. When users gave negative comments or suggestions, the evaluator often considered them to be usability problems. But, of course, not all negative comments or suggestions are problems. For example, in the Danish pairs' tests, the total number of the negative comments in the testing was 167, much larger than the number of negative comments in the problem period 99 (97+2 including the negative comments in both UPU and UPE period). This showed that evaluators did not take users' every negative comment as a problem, and they may have used their own judgment in deciding which negative comment was related to the problem and which was not. However, since the total rates per minute of the negative comments and suggestions were significantly higher than those in the selected videos, it indicates that negative comments and suggestions were given more often in the UPU period than in the other period. The evaluators may have taken the users' negative comments and suggestions as important clues to learning the users' problems.

On the other hand, in the UPE period, the users more often gave a "short reply" than negative comments or suggestions. The UPEs are from the evaluators' pre-thoughts and the users mainly agree or disagree with the evaluators' comments, and so a "short reply" was often used. Thus, we can see that in the UPU period, the user's main role is the sender, whereas in the UPE period, the user's main role is the receiver.

The users' point event behaviours were similar in all cultural settings in the problem period, indicating that users with different cultural backgrounds tended to behave similarly when there were usability problems. This finding suggests that regardless of which cultural backgrounds the users have, usability professionals may take the users' specific communication patterns as signals for identifying usability problems.

Theme 4d: Evaluators' communication patterns for the UPU and UPE are different in different cultural settings.

The findings in theme 4d were: 1) In the UPU period, the digging deeper probes were significantly more than the other probes; 2) Evaluators gave more digging deeper probes per minute in the UPU period than in the UPE period, whereas they gave more directed probes per minute in the UPE period than in the UPU period; 3) East Asian evaluators were more willing to provide help than were Western evaluators, especially when they were with East Asian users.

In the UPU period, the evaluators gave significantly more digging deeper probes than other probes. The finding indicates that when the problem was learnt from the users, the evaluators often used digging deeper probes to make the users explain more about the problem in order to find the right problem and to improve it in the future. There were few directed probes when the problem was from the users. The directed probing was the question or topic started by the evaluators, which was not closely related to the user's current talk, behaviour or emotion. The findings imply that normally the evaluators followed with the users' speech, behaviour or emotion when the problem was from the users' experience.

However, in the UPE period, the evaluators gave more directed probes than digging deeper probes. The result indicates that if the problem was already in the evaluators' mind, it wasn't necessary to use a lot of questions to ask the users to explain or clarify; instead, the evaluators had to point the issue out and get the feedback from the users.

As in the testing period, in the problem period, East Asian evaluators tended to give more help than did Western evaluators. Moreover, East Asian evaluators' help to local users occurred more often compared to that of foreign users. The reasons might be: 1) In order to get the foreign users' real problems, the evaluators tended to give less help. 2) It may be easier to provide help to local users than to foreign users.

Summary of theme 4

For the communications in the problem period, we can see that there were different communication patterns for the UPU and UPE, whereas the culture's impact seemed limited. According to Ericsson and Simon's thinking aloud model, and Boren and Ramey's thinking aloud model, the problem information should be transmitted from users to evaluators. In contrast, this research indicates that in most situations the problem information was transmitted from users to evaluators, and sometimes the information was also transmitted from evaluators to users. Western evaluators tended to be more active than were East Asian evaluators in transmitting problem information to users. Moreover, when the problem information was transmitted from users to evaluators, compared to local evaluators, foreign evaluators relied more on the objective measure "time" to rate the severity. Furthermore, users' state events had similar tendencies as the findings in the testing period, such as spending more time on "talking rather than thinking aloud" compared to thinking aloud and silence, and Western users were better at thinking aloud than were East Asian users. Finally, there were many differences for the users' and evaluators' point event behaviours in the UPU and UPE period, but there were few differences in different cultural settings. Culture seems not to play an important role for the point event behaviours in the problem period.

5.4 Evaluators' View on the Thinking Aloud Usability Testing

After the tests, the evaluators were interviewed about their criteria of deciding the usability problems, the severity of the problems and their view of the culture's impact on usability testing.

From the interviews, we recognized that usability problems were identified by considering four aspects: users' verbal and non-verbal behaviors (including the users' task performing behaviors, comments and suggestions, the users' expectations and the frequency of the problem), clients' requirements, the purpose of the application and the evaluators' own experience.

In the interviews, all evaluators mentioned the users' verbal and non-verbal behaviors. For the seven criteria listed in section 4.5.2, four of these were related to users. Usability testing is a user involved UEM in which the benefit is to provide direct information about how users use the application and what their exact problems are (Nielsen, 1993), and thus the users' task performing and other verbal and non-verbal behaviors are important in identifying problems.

However, from the interviews, we found that when the evaluators were trying to identify the problems, apart from considering the users' verbal and non-verbal behaviors, some evaluators also considered the clients' requirements, the purpose of the application and even their own experience. This finding shows usability testing might not be as objective as previous researchers thought (Hartson et al., 2001; Hvannberg et al., 2007; Yeo, 2001b). When users met problems or gave negative comments, the evaluators may have thought of several other issues to decide on whether they were usability problems or not. For example, if from the evaluators' perspective, it was not a problem, even though the user mentioned it, the evaluator may have ignored it, unless it was mentioned by several users.

Regarding the judgment of the problem severity, apart from the above four aspects, the evaluators also mentioned "the influence of the redesign." For example, an evaluator said, "If the problem is easy to fix, I would rate it as a minor problem. But if it seems hard or expensive to fix, I would rate it as important or critical."

The Danish and Chinese evaluators' criteria of deciding on the usability problems and the problem severities were similar, indicating the role that the evaluators' professional knowledge played in finding and rating the problems. However, when the evaluators conducted the usability testing with users in another country, the criteria may have changed. For example, the evaluators in the interviews reported that they would rely less on their own experience when doing tests in another country.

With respect to interviews of the culture's impact on the usability testing, most evaluators regarded language as being the most important issue. They thought that if people could share the same language, there would be no problem in doing tests abroad. However, apart from the language issues, some evaluators also mentioned that it was easier for them to do tests with local users. For example, Chinese evaluators said that it was not necessary for Chinese users to express everything and they could still understand, but with foreign users, the Chinese evaluators would expect to get all the information from the users. Further, the evaluators also agreed that Chinese users were less talkative than were Danish users, which implied that it might be more difficult for the foreign evaluators to gain access to Chinese users' thoughts. For the tests in this study, gender and gestures were not considered by evaluators. However, some evaluators said that gender and gestures might need to be considered in some cultures, such as Indian or Muslem cultures.

5.5 Chapter Summary

This chapter has focused on discussing the findings in this study. The contribution and limitation of this research will be discussed in the conclusion chapter.

The research investigates Danish and Chinese evaluators' and users' behaviors at the group level and the focus of the research is the cultural variation approaches to usability testing. The demographic information shows that the users and evaluators were similar in different cultural settings and were able to represent the target people. However, if the findings in different cultural settings were different, it could implicate the impact of the evaluators' and users' cultural backgrounds on usability testing.

This chapter has discussed the usability problem finding and problem severity rating behaviors by considering the culture theories in this research and other people's studies. Compared to problem finding behaviors, problem severity rating behaviors seem more sensitive to the evaluators' and users' cultural backgrounds. Moreover, the cultural impact on thinking aloud usability testing was not the same for tests with Westerners and East Asians. It may not have been an important issue for foreign evaluators to do tests with Western users who tended to have analytic cognitive style and low-contextual communication orientation. However, it may have been more difficult for foreign evaluators to do tests with East Asians users who tended to have a holistic cognitive style and high-contextual communication.

The investigation of the communication patterns was an explorative study, and thus themes rather than hypotheses were examined. Culture plays an important role in the users' and evaluators' communication patterns in the testing. However, for the communication patterns in the problem period, culture's impact seems limited. Instead, the communication patterns in the UPU and UPE period were quite different.

Finally, the interviews with the evaluators were discussed in order to understand the evaluators' problem finding and problem severity rating behaviors and their view of the culture's impact on usability testing. From the discussion of the criteria of deciding problems and problem severities, we can see that beyond the "objective" criteria, such as the users' verbal and non-verbal behaviors, the evaluators also used some "subjective" criteria, such as the evaluators' own experience, which might have been influenced by culture. Furthermore, culture's impact on the process of the usability testing from the evaluators' perspective was also briefly discussed.

The conclusion chapter follows next.

6 Conclusion

In this conclusion chapter, we revisit the research question proposed in the introduction chapter and review the major findings of this dissertation. Then, the implication of this research is discussed from the theoretical, methodological and practical perspectives. Finally, limitations of this research and recommendations for future research are identified.

6.1 Revisiting the Research Question and Summarizing the Major Findings

The research question in this dissertation is: *To what extent does the evaluators' and users' cultural background influence the thinking aloud usability testing?*

In order to investigate this research question, evaluators and users with the same and different cultural backgrounds participated in this study. Nisbett's analytic- and holistic- cognitive styles and Hall's low-contextual and high-contextual communication theories are the main concerns for the concept of "culture" in this research, and so the "cultural backgrounds" in the research question should hold the cognitive styles and communication orientations as the culture theories described. Danish people are regarded as having an analytic cognitive style and low-contextual communication, and Chinese people are regarded as having a holistic cognitive style and high-contextual communication (Hall, 1989a, 1990; Hall & Hall, 1990; Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005; Nisbett & Norenzayan, 2002b). The research question has been investigated through a comparison of the thinking aloud usability tests conducted by Danish and Chinese evaluators and users. The tests with Danish users were held in Denmark, and the tests with Chinese users in China. The findings in this research give implications for the thinking aloud usability testing in Western and East Asian countries.

The thinking aloud usability testing has been investigated from two perspectives: testing result and testing process. The purpose of the usability testing is to find usability problems (Nielsen, 1993), and thus usability problems can be regarded as the testing result. In order to find the problems, communication between the users and evaluators is very important (Boren & Ramey, 2000; Shi & Clemmensen, 2008). The communication can be regarded as the testing process. The identified usability problems and communications for the evaluators and users with different cultural backgrounds have been examined.

The research question is answered from the testing result and testing process perspectives. The major findings are shown in Table 55.

Table 55: Major findings in this research

Major findings
Usability testing result (Usability problems)
1. Evaluators' cultural backgrounds seem unimportant for the usability problem identification in the thinking aloud usability tests. Foreign usability specialists are able to find similar amount and similar types of usability problems as the local usability specialists when there are sufficient test sessions, especially for the tests with Western users. (see section 4.2.1, section 4.3.1 and section 5.2.1)
2. Evaluators' and users' cultural backgrounds influence the usability problem severity rating, as predicted by the culture theories. East Asian evaluators tend to give the middle severity rank to the problems, whereas Western evaluators do not have such tendency. (see section 4.2.2 and section 5.2.2.1)
3. Culture's influence on thinking aloud usability testing is not the same in the Western and East Asian settings. For the tests with Western users who tend to have analytic cognitive style and low-contextual communication orientation, evaluators with different cultural backgrounds may be able to do the usability tests with similar results (usability problem identification and problem severity rating). However, for the tests with East Asian users who tend to hold holistic cognitive style and high-contextual communication orientation, evaluators' cultural backgrounds may have some influence on the result of the thinking aloud usability testing. (see section 4,2,1, section 4.3.2, section 5.2.1.1 and section 5.2.2.2)
Usability testing process (Communications)
1. Users rarely think aloud in the sense of Ericsson and Simon (1993) described. Western users do more thinking aloud than do East Asian users, and East Asian users are more willing to be silent than are Western users. (see section 4.4.2 and section 5.3.1.1)
2. Apart from differences in thinking aloud, users' verbal behaviours show some differences in different cultural settings. Western users give more negative comments than do East Asian users. Users, especially East Asian users, give more culture related comments to foreign evaluators than they do to local evaluators. (see section 4.4.3 and section 5.3.1.2)
3. Evaluators' verbal behaviours are different in different cultural settings. Western evaluators are more active than East Asian evaluators in probing questions and giving affirmative responses and reminders, whereas East Asian evaluators are more active than Western evaluators in providing help. Moreover, users' cultural backgrounds have influence on the evaluators' behaviours. All evaluators tended to give more reminders to East Asian users than to Western users. (see section 4.4.4 and section 5.3.1.3)
4. UPU and UPE which are put forward based on the direction of the usability problem information transmission do involve different communication patterns in all cultural settings. Western evaluators

have more UPEs than do East Asian evaluators. Foreign evaluators may rely more on the objective measure “time” to rate the severity of the UPU than do local evaluators. (see section 4.4.5 and section 5.3.1.4)

From table 55, we can see that the evaluators’ cultural backgrounds do have some influences on the thinking aloud usability testing result. However, the influence is different for the tests with Western and East Asian users. For the thinking aloud usability testing process, evaluators and users with different cultural backgrounds tend to have different communication patterns in different cultural settings. In the next section, the implications of this research are discussed.

6.2 Implications of this Research

The present research investigates the evaluators’ and users’ cultural backgrounds on the thinking aloud usability testing. The research has theoretical, methodological, and practical implications for both researchers and practitioners, as discussed below.

6.2.1 Theoretical Implications

The dissertation has at least six implications for usability research.

First, this study indicates that Boren and Ramey’s thinking aloud model (2000) is more suitable for thinking aloud usability testing than is Ericsson and Simon’s classic model (1993). Both Danish and Chinese users seldom followed the classic thinking aloud model. The classic model is an unnatural way to talk. Thinking aloud usability testing is often used in formative evaluation, the purpose of which is to find usability problems, and as long as the relaxed talk does not cover the problems, it is reasonable to do so. From this research, we can see that users preferred the relaxed talk and evaluators always communicated with the users during the tests. We conclude thus that Boren and Ramey’s thinking aloud model is more appropriate for usability testing.

However, the findings in this research actually go beyond Boren and Ramey’s thinking aloud model (2000). In their model, the evaluators are the learners who learn the usability problems from the users. The authors suggest that evaluators play only a small role as technique expert to troubleshoot the application, and that this role should not be emphasized. The findings in this study show that the evaluators sometimes played a role in initiating usability problems. Moreover, when the users gave negative comments or suggestions, the evaluators would use their own minds to decide whether those were problems or not. This research thus implicates that in usability testing, it is not the user alone who plays the role in troubleshooting the usability problems; the interaction between the users and evaluators may need to have more emphasis for deciding on the problems and problem severities. The findings imply that usability testing cannot be treated as a "gold standard" against which to compare other UEMs (Hartson et al., 2001; Hornbæk, 2009; Hvannberg et al., 2007; Yeo, 2001b).

Second, Nisbett and his colleagues' findings of cognitive styles and Hall's theory of communication orientations are applied to the empirical study in order to investigate the cultural issues in usability testing. Many previous cultural or cross-cultural usability studies have been based on Hofstede's culture theories (Callahan, 2005; Marcus, 2005; Marcus & Gould, 2000; Vatrapu & Pérez-Quiñones, 2006; Yeo, 1998a, 2001b). This research indicates a relatively new angle to investigating cultural issues in UEMs, that is, a more psychological and more process orientated perspective.

Third, this research finds evaluators' and users' cultural backgrounds do impact the usability testing, but the influence is not the same for participants with different cultural backgrounds. The research implies that users' different cognitive styles and communication orientations may have different impacts on the thinking aloud usability testing result and the usability testing process.

Fourth, the research indicates that culture's impact on usability problem finding behaviours is limited, but severity rating behaviours and communications are more sensitive to the evaluators' and users' cultural backgrounds.

Fifth, from this dissertation, we get a better understanding of the communications in the usability testing. There are limited studies about communications in usability testing. This research investigates Danish and Chinese evaluators' and users' communication patterns systematically. The findings indicate that Western and East Asian evaluators and users tend to have different communication patterns in usability testing. However, for communication patterns in the problem periods, there were not so many differences in different cultural settings. Instead, the problem communication patterns vary in different ways of transmitting problem information, i.e. evaluators and users have different communication patterns for the UPU and UPE.

Sixth, this research identifies some explanations for other people's studies: 1) The dissertation indicates that the interaction between users and evaluators may play an important role in finding and rating usability problems, which may help us gain a better understanding of the evaluator effect (Hertzum & Jacobsen, 2001) (see section 2.3.1.1 and section 5.2.1.1); 2) A study done by Hassenzahl (2000) showed a lack of correspondence between the time on the problem and the severity of the problem given by evaluators. However, this research indicates actually there is a relation between the time on the problem and the problem severities if considering the evaluators' familiarity of the problems. This research found foreign evaluators relied more on the objective measure "time" to rate the severity of the UPU, implying that when the problem is learnt from users and evaluators are not familiar with the target user group, the evaluators rely more on the objective measure to decide on the problem severity (section 2.3.1.1 and section 5.3.1.4).

6.2.2 Methodology Implications

This research indicates that the experimental study, used for this research, is suitable for investigating thinking aloud usability testing. The reasons are:

1) Thinking aloud usability testing is a laboratory testing, and thus using the experiment to investigate usability testing will not bring much more artificial and unnatural factors than the usability testing itself.
2) With the comparable studies in different cultural settings, we gained a deeper understanding of the impact of evaluators' and users' cultural backgrounds on thinking aloud usability testing.
3) The researcher could be regarded as the client in this study. The evaluators were usability specialists, through which better external validity guaranteed compared to using students or researchers, which has been done in a number of earlier studies (Hornbæk & Frøkjær, 2005; Krahmer & Ummelen, 2004a; Vatrapu & Pérez-Quiñones, 2006). The equipment and testing rooms were exactly the same as those of usability testing in practice. The researcher gave all evaluators the same instructions for the tests and application. The evaluators were asked to do the usability testing as they usually did. Therefore, the findings in this research should be able to be generalized as the usability testing in practice, at least to some extent.

Moreover, in this research, apart from comparing the number of the problems in different cultural settings, the types of the problems were also considered and analyzed. Many previous studies mainly have focused on comparing the number of problems in order to assess different UEMs or different conditions of one UEM (Als et al., 2005; Hvannberg et al., 2007; Law &

Hvanneberg, 2004; Virzi et al., 1996). Some researchers have criticized this approach (Hartson et al., 2001; Hornbæk, 2009; Keenan et al., 1999), and suggested classifying the problems by types (Hartson et al., 2001). This study categorized the problem instances given by the evaluators into 30 non-overlapping problem types. Since matching problem description is not a straightforward activity (Hornbæk, 2009), the research also took into account the way of categorizing the problem instances (section 3.5.1.1 and section 3.5.1.2).

Further, we developed a coding system based on the thinking aloud models, culture theories and previous studies, and analyzed the communication patterns systematically in this research. There are very few studies that thoroughly analyze the communication in usability testing. Previous research has mainly discussed the evaluators' and users' communication patterns qualitatively, not quantitatively (Boren & Ramey, 2000; Nørgaard & Hornbæk, 2006; Shi, 2008a). From previous studies, we have known that some communication patterns may exist in usability testing, but we have not known which verbal behaviours are more important in usability testing and for finding usability problems. We are also not clear whether or not these communication patterns are the same for participants with different cultural backgrounds. Through the systematic coding and statistical analysis, we have gained a clearer view of the evaluators' and users' communication patterns in usability testing in different cultural settings.

Finally, in this research, apart from the experiment itself, interviews were also conducted to gain more information from the users and evaluators. The data from the interviews has enriched the quantitative analysis from the experiment.

6.2.3 Practical Implications

The findings in this research have eight implications for usability practitioners in Western and East Asian cultures.

First, East Asians are not good at thinking aloud, so evaluators' reminders and probing behaviors may be more important for the tests with East Asian users. When with East Asian users, it may be harder to follow Ericsson and Simon's classic thinking aloud model. East Asian users tend to do the task silently and then talk their thoughts, or talk their thoughts first and then do the task silently. Their preferred thinking aloud model is not talking and speaking at the same time, but more like "doing silently- talking- doing silently- talking". In order to get the East

Asian users' usability problems, observing the users' task performing behaviors and listening to their speech are equally important for the evaluators.

Second, East Asian users tend to be high-contextual communication and sometimes they may not talk everything since some information is supposed to be known by the evaluators. So when doing tests with East Asian users, the evaluators may need to give many digging deeper probes to get the users' real feelings and real problems.

Third, compared to the tests with East Asian users who tend to have holistic cognitive style and high-contextual communication orientation, it may be easier for foreign evaluators to do tests with Western users who tend to be better at thinking aloud and with low-contextual communication.

Fourth, East Asian evaluators may have the tendency to give the middle rank to the usability problems when rating the severities. So it may be better for the severity scales to have even number, not an odd number, of ranks.

Fifth, East Asian users may not be as critical as are Western users, but it does not mean that they are satisfied with the application. East Asian users may not be willing to give many negative comments, though they have problems. Usability practitioners should be aware of the East Asians' socio-emotional orientation and identify the problems based on both their comments and their behaviors.

Sixth, evaluators' own experience may play a role in finding and rating the usability problems. However, for tests with foreign users, usability practitioners should rely less on their own feelings, but more on the users' task performing behaviours and comments.

Seventh, compared to the usability problem finding behaviours, the findings show that problem severity rating behaviours are more sensitive to the evaluators' cultural backgrounds. This implies that it may not be a problem for usability practitioners to do tests in different cultures if finding usability problems is the main purpose. However, if rating the problem severity is also very important, then it may be better to use local evaluators.

Eighth, communication seems important for usability tests in both Western and East Asian cultures. In order to find usability problems, usability practitioners may need to pay attention to users' negative comments, suggestions and questions. In order to encourage the users to talk, the evaluators need to give lots of affirmative responses. In order to understand the users' problems, the evaluators need to give digging deeper probes. For the potential problems that the users do not notice, evaluators may need to direct the users' attention on those issues and to get feedback from the users.

6.3 Limitations of this Research

This thesis has some limitations which could be improved in future work.

First, there are some factors which may threaten the external validity. External validity refers to the generalization of the findings in one study across populations, settings or time (Maxwell & Delaney, 2004). In this study, all participants were required to speak English in the tests. However, speaking English in a usability testing situation may be different from testing in practice. From the field study (Clemmensen et al., 2007), we know that Danish and Chinese usability practitioners normally speak their local language when doing tests with local users, but usually speak English when doing tests with foreign users. There have been many studies carried out on local pairs' usability testing in specific cultures (Clemmensen & Shi, 2008; Clemmensen et al., 2007; Lee & Lee, 2007; Shi, 2008a; Yeo, 1998a), but there are limited studies which compare local and foreign evaluators' usability testing. In the interviews in this study, most evaluators reported that the major difference for the tests with local and foreign users was language. Many of them thought that if people could speak the same language, there would be no other differences. Thus, in order to see factors beyond the influence of language in this research, that is, to see whether there are any other differences between the local and distant pairs' usability testing, it was better to ask all participants to speak the same language. Additionally, speaking the same language would show better internal validity, thus ensuring that the difference between the local pairs and distant pairs was from the participants' different cultural backgrounds (i.e., cognitive styles and communication orientations), and not the language used in the tests. Furthermore, the researcher did not understand Danish. If asking the Danish pairs to speak Danish in the tests, there would have also been translation issues (Dray & Siegel, 2005). Therefore, in this research, asking the participants to speak English seemed a better choice, but with the limitation that the generalization of the findings for the local pairs' testing may be limited. Future studies could explore the impact of languages on the thinking aloud usability testing to compare the differences between speaking local languages and English.

A second limitation is that the participants were those who were good at English, and thus the users and evaluators may represent a small population in Denmark and China. However, since all participants were local people who were born and raised in Denmark/ China, they could, to some extent, be considered as being representative.

Third, the findings in this research are supposed to be generalized to the thinking aloud usability testing in Western and East Asian countries. However, the study involved only two countries. In future, more countries should be involved.

Fourth, the evaluators were asked to write down the usability problems immediately after each test, which means that they came up with problems based on their notes in the tests, not on the videos of the tests. This could be questioned, since it could be argued that it might be different from usability testing in practice. However, in the field study (Clemmensen et al., 2007), we have found that in practice, even though most usability tests use video recording, very few usability practitioners have come up with usability problems based on the videos after the tests. Instead, they prefer to take notes and come up with the problems during the test or immediately after each test. “Discount usability engineering” (Nielsen, 1993, p. 17) seems popular in practice now. However, the way that usability problems were brought forward in this study may, to some extent, limit the research findings.

In addition to the limitations of external validity, the threats to internal validity also need to be clarified. Internal validity refers to whether the measurements are due to the manipulation (Field & Hole, 2003). Through selecting similar users and evaluators, asking the participants to speak the same language, and giving the same instructions to participants, a good internal validity should be guaranteed. However, this research may still have a threat to the internal validity. Nisbett and his colleagues’ studies about analytic and holistic cognitive styles and Hall’s low- and high- contextual communication orientations are the two main culture theories related to this research. From their studies and theories, we learn that Danish and Chinese people have different cognitive styles and communication orientations. However, in this research, we did not really test whether the participants had the cognitive styles and communication orientations as the theories described. We had only two questions as detectors to check the participants’ cultural orientations, which was far from enough. In future work, it would be better to develop a convenient tool to examine the participants’ cultural orientations thoroughly. However, since the focus of this research is not to examine Nisbett’s and Hall’s theories, but to apply their theories in usability testing settings, the Danish and Chinese participants who were selected from the target countries could have had different tendencies as the theories described, at least to some extent (see section 3.1.2.1, section 3.2.1 and section 5.1).

This section discusses the limitations in this research. The external and internal validity may need to be improved in future work. In the next section, recommendations for further research are discussed.

6.4 Recommendations for Further Research

This thesis provides empirical evidence concerning the differences of thinking aloud usability testing conducted in different cultures with local and foreign evaluators. From this research, we have gained insights for usability research and usability practice on how to do usability testing efficiently in different cultures. Future research should examine whether usability testing could be improved based on the advice given in this study.

Also, the culture theories in this research are based mainly on Nisbett and his colleagues' studies and Hall's work. As discussed in section 5.2.1.2, culture's impact may not be static, but related to the situation. Hong and Mallorie's (2004) dynamic constructivist approach to culture could be considered in the future to investigate when and in which conditions the cultural differences appear or disappear in the usability testing setting.

Further, with the advent of the product and systems internationalization, users of a specific application are becoming more and more similar in different cultures. Although usability testing is a relatively new technique which originated in Western countries, it is now used all over the world. In usability tests, if testing globalized application with international users, there may be fewer differences between local and foreign evaluators than in testing localized applications. Future research could compare usability tests with globalized and localized applications in different cultures to see whether or not the evaluators' cultural backgrounds have a similar impact.

Moreover, culture is a broad concept with various levels, such as professional culture, organizational culture, national culture, technical culture vs. user-centered culture, etc. (Beu et al., 2000; Bloor & Dawson, 1994; Iivari, 2004, 2006; Iivari & Abrahamsson, 2002). Culture in this research refers mainly to the common understanding of culture which is nationally orientated. Further research could also consider other cultural levels, such as organizational culture.

As a penultimate recommendation, future research could carry out the same tests but ask the local pairs to speak local languages, and then the findings in the two studies could be compared. Finally, apart from thinking aloud usability testing, culture's impact on other UEMs would also be worthwhile investigating.

7 References

Als, B. S., Jensen, J. J., & Skov, M. B. (2005). Comparison of think-aloud and constructive interaction in usability testing with children. Paper presented at the Proceedings of the 2005 conference on Interaction design and children. Pages 9-16.

Andersen, K. G. (2001). Communicating culture. Denmark: Aalborg University printing house.

Andre, T. S., Hartson, H. R., Belz, S. M., & McCreary, F. A. (2001). The user action framework: a reliable foundation for usability engineering support tools. International Journal of Human Computer Studies, 54(1), 107-136.

Aryee, S., Luk, V., & Fields, D. (1999). A cross-cultural test of a model of the work-family interface. Journal of Management, 25(4), 491.

Barnum, C. M. (2002). Usability testing and research: Longman.

Barrett, P. (2001). Assessing the reliability of rating data. Retrieved January21, 2009 from http://www.pbarrett.net/techpapers/rater.pdf.

Benbunan-Fich, R. (2001). Using protocol analysis to evaluate the usability of a commercial web site. Information & Management, 39(2), 151-163.

Bennett, M. J. (1998). Intercultural communication: A current perspective. Basic concepts of intercultural communication: Selected readings, 1-34.

Beu, A., Honold, P., & Yuan, X. (2000). How to build up an infrastructure for intercultural usability engineering. International Journal of Human-Computer Interaction, 12(3/4), 347.

Bilal, D., & Bachir, I. (2007). Children's interaction with cross-cultural and multilingual digital libraries. II. Information seeking, success, and affective experience. Information Processing & Management, 43(1), 65.

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. Paper presented at the CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems. P.31-40.

Bloor, G., & Dawson, P. (1994). Understanding professional culture in organizational context. Organization Studies, 15(2), 275.

Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication, 43(3), 261-278.

Bourges-Waldegg, P., & Scrivener, S. A. R. (1998). Meaning, the central issue in cross-cultural HCI design. Interacting with Computers, 9(3), 287-309.

Bourges-Waldegg, P., & Scrivener, S. A. R. (2000). Applying and testing an approach to design for culturally diverse user groups. Interacting with Computers, 13(2), 111.

Briley, D. A., Morris, M. W., & Simonson, I. (2000). Reasons as carriers of culture: Dynamic versus dispositional models of cultural influence on decision making. Journal of consumer research, 27, 157-178.

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (eds.) Usability Evaluation in Industry London: Taylor and Francis., http://www.usabilitynet.org/trump/documents/Suschapt.doc.

Bryman, A. (2008). Social research methods (third ed.). New York: Oxford University Press Inc.

Bryman, A., & Bell, E. (2007). Business research methods. New York: Oxford University Press Inc.

Bull, S., & McCalla, G. (2002). Modelling cognitive style in a peer help network. Instructional Science 30(6).

Callahan, E. (2004). Interface design and culture. In B. Cronin (Ed.), Annual Review of Information Science and Technology (pp. 257-310).

Callahan, E. (2005). Cultural similarities and differences in the design of university websites Journal of Computer-Mediated Communication, 11(1), article 12. http://jcmc.indiana.edu/vol11/issue11/callahan.html.

Capra, M. G. (2006). Usability problem description and the evaluator effect in usability testing. Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Carey, J. M. (1998). Creating global software: A conspectus and review. Interacting with Computers, 9(4), 449.

Choi, I., & Nisbett, R. E. (2000). Cultural psychology of surprise: Holistic theories and recognition of contradiction. Journal Of Personality And Social Psychology, 79(6), 890-905.

Choong, Y.-Y. (1996). Design of computer interfaces for the Chinese population.

Choong, Y.-Y., & Salvendy, G. (1999). Implications for design of computer interfaces for Chinese users in Mainland China. International Journal of Human-Computer Interaction, 11(1), 29.

Cinnirella, M., & Green, B. (2007). Does `cyber-conformity' vary cross-culturally? Exploring the effect of culture and communication medium on social conformity. Computers in Human Behavior, 23(4), 2011.

Clark, H. H. (1991). Communities, commonalities, and communication. In J. J. Gumperz & S. C. Levinson (Eds.), Rethinking linguistic relativity (pp. 324-355): Cambridge University Press.

Clemmensen, T. (2006, May 14-15). Cultural models in psychological usabililty evaluation methods (UEM). Paper presented at the Indo-Danish HCI Research Symposium Guwahati, India.

Clemmensen, T., & Goyal, S. (2005). Cross cultural usability testing. Working paper, Copenhagen Business School, Department of Informatics, HCI research group, 2005-006.

Clemmensen, T., Hertzum, M., Hornbaek, K., Shi, Q., & Yammiyavar, P. (2008). Cultural cognition in the thinking-aloud method for usability evaluation. Paper presented at the ICIS 2008 Proceedings. Paper 189., Paris.

Clemmensen, T., Hertzum, M., Hornbaek, K., Shi, Q., & Yammiyavar, P. (2009). Cultural cognition in usability evaluation. Interacting with Computers, 21, 212-220.

Clemmensen, T., & Shi, Q. (2008). What is part of a usability test? . Paper presented at the CHI 2008, Florence. P. 3603-3608.

Clemmensen, T., Shi, Q., Kumar, J., Li, H., Sun, X., & Yammiyavar, P. (2007). Cultural usability tests - How usability tests are not the same all over the world. Paper presented at the HCI International Beijing, China. P.281-290 (Lecture Notes on Computer Science; 4559).

Clemmensen, T., Shi, Q., Sun, X., & Yammiyavar, P. (2008). Localizing usability evaluation methods: why use local moderators in think aloud usability tests? Unpublished Working Paper.

Clemmensen, T., Yammiyavar, P., & Sun, X. (2006). Reminders, gestures and languages in the think-aloud UEM in cross cultural usability testing.Unpublished manuscript.

Cole, M. (1996). Interacting minds in a life-span perspective: A cultural/historical approach to culture and cognitive development. Interactive minds: Life-span perspectives on the social foundation of cognition, 59-87.

Colman, A. M. (2001). A dictionary of psychology retrieved from "http://www.encyclopedia.com/doc/1O87-cognitivestyle.html ".

Coolican, H. (2004). Research methods and statistics in psychology (Fourth ed.). London.

Cozby, P. C. (2003). Methods in behavioral research (eighth edition ed.): McGraw-Hill.

Creswell, J. W. (2003). Research design - qualitative, quantitative and mixed method approaches. London: SAGE.

Day, D. L. (1998). Shared values and shared interfaces: The role of culture in the globalisation of human-computer systems. Interacting with Computers, 9(3), 269.

De Angeli, A., Athavankar, U., Joshi, A., Coventry, L., & Johnson, G. I. (2004). Introducing ATMs in India: a contextual inquiry. Interacting with Computers, 16(1), 29.

Downey, S., Wentling, R. M., Wentling, T., & Wadsworth, A. (2005). The relationship between national culture and the usability of an E-learning System. Human Resource Development International, 8(1), 47.

Dray, S. M., & Siegel, D. A. (2005). "Sunday in Shanghai, Monday in Madrid" Key issues and desisions in planning international user studies. In N. Aykin (Ed.), Usabillity and Internationalization of Information Technology (pp. 189-212). London: Lawrence Erlbaum.

Dumas, J. S., & Loring, B. A. (2008). Moderating usability tests- principles and practices for interacting. Burlington: Morgan Kaufmann Publishers, Elsevier Inc.

Dumas, J. S., & Redish, J. C. (1999). A practical guide to usability testing (Revised Edition ed.). Portland, Oregon, USA: Intellect.

Duncker, E. (2002). Cross-cultural usability of the library metaphor. Paper presented at the Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries Portland, Oregon, USA. P.223 - 230

Eide, H., Quera, V., & Finset, A. (2003). Exploring rare patient behaviour with sequential analysis: An illustration. Epidemiologia e Psichiatria Sociale, 12, 109-114.

Ericsson, K. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. Mind, Culture, & Activity, 5(3), 178-186.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data: Cambridge Massachusetts.

Faiola, A., & Matei, S. A. (2005). Cultural cognitive style and web design: Beyond a behavioral inquiry into computer-mediated communication. Journal of Computer-Mediated Communication, 11(1).

Fang, X., & Rau, P.-L. P. (2003). Culture differences in design of portal sites. Ergonomics, 46(1-3), 242.

Field, A. (2005). Discovering statistics using SPSS. London, Thousand Oaks, New Delhi: SAGE Publications.

Field, A., & Hole, G. (2003). How to design and report experiments. London: SAGE Publications Inc.

Fisher, C., & Sanderson, P. (1996). Exploratory sequential data analysis: Exploring continuous observational data. Interactions, 25-34.

Frandsen-Thorlacius, O., Hornbæk, K., Hertzum, M., & Clemmensen, T. (2009). Non-universal usability?: a survey of how usability is understood by Chinese and Danish users. Paper presented at the CHI '09: Proceedings of the 27th international conference on Human factors in computing systems. P.41-50.

Galdo, E. M. d., & Nielsen, J. (1996). International user interfaces. New York, Chichester, Brisbane, Toronto, Singapore: John Wiley & Sons, Inc.

Gillan, D. J., & Bias, R. G. (2001). Usability science. I: Foundations. International Journal of Human-Computer Interaction, 13(4), 351-372.

Graugaard, P. K., Holgersen, K., Eide, H., & Finset, A. (2005). Changes in physician–patient communication from initial to return visits: a prospective study in a haematology outpatient clinic. Patient Education and Counseling, 57(1), 22-29.

Gray, W. D., & Salzman, M. C. (1998a). Damaged merchandise? A review of experiments that compare usability evaluation methods. Human-Computer Interaction, 13(3), 203-261.

Gray, W. D., & Salzman, M. C. (1998b). Damaged merchandise? A review of experiments that compare usability evaluation methods. . Human-Computer Interaction, 13(3), 203-261.

Griffith, T. L. (1998). Cross-cultural and cognitive issues in the implementation of new technology: focus on group support systems and Bulgaria. Interacting with Computers, 9(4), 431.

Guan, Z., Lee, S., Elisabeth, C., & Judith, R. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. Paper presented at the CHI. P.285-294.

Gumperz, J. J., & Levinson, S. C. (1991). Rethinking linguistic relativity: Cambridge University Press.

Gunykunst, W. B. (1993). Communication in Japan and the United States. New York: State University of New York Press.

Gutierrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. Educational Researcher, 32(5), 19-25.

Hall, E. T. (1989a). Beyond culture. New York: Anchor Books/ Doubleday.

Hall, E. T. (1989b). The dance of life--the other dimension of time. New York, London, Toronto, Sydney, Auckland: Anchor Book/ Doubleday.

Hall, E. T. (1990). The silent language. New York: Anchor Books, a division of Random House, Inc.

Hall, E. T., & Hall, M. R. (1990). Understanding cultural differences. Yarmouth, Maine: Intercultural press, INC.

Hall, M., De Jong, M., & Steehouder, M. e. (2004). Cultural differences and usability evaluation: Individualistic and collectivistic participants compared. Technical Communication, 51(4), 489.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction, 13(4), 373-410.

Hassenzahl, M. (2000). Prioritizing usability problems: data-driven and judgement-driven severity estimates. Behaviour & Information Technology, 19(1), 29-42.

Hayes, J., & Allinson, C. W. (1996). The implications of learning styles for training and development: a discussion of the matching hypothesis. British Journal of Management, 7(1), 63-73.

Hayes, S. C. (1986). The case of the silent dog-Verbal reports and the analysis of rules: A review of Ericsson and Simon's protocol analysis: Verbal reports as data. Journal of the Experimental Analysis of Behavior, 45(3), 351-363.

Heineman, P. L. (retrieved in Jan. 2008). Field dependent-independent research. http://www.personality-project.org/perproj/others/heineman/eft.htm.

Hertzum, M. (2006). Problem prioritization in usability evaluation: from severity assessments toward impact on design. International Journal of Human-Computer Interaction, 21(2), 125-146.

Hertzum, M., Clemmensen, T., Hornbæk, K., Kumar, J., Shi, Q., & Yammiyavar, P. (2007). Usability constructs : A cross-cultural study of how users and developers experience their use of information systems. Paper presented at the HCI International, Beijing, China.

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

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction, 13(4), 421-443.

Hinton, P. R., Brownlow, C., McMurray, I., & Cozens, B. (2004). SPSS Explained. New York: Routledge Inc.

Hofstede, G. H. (2001). Culture's consequences: comparing values, behaviors, institutions, and organizations across nations (2 ed.). Thousand Oaks, London, New Delhi: Sage Publications- International educational and professional publisher.

Hong, Y.-y., & Mallorie, L. M. (2004). A dynamic constructivist approach to culture: Lessons learned from personality psychology. Journal of Research in Personality, 38, 59–67.

Honold, P. (2000). Cultural and context: an empirical study for the development of a framework for the elicitation of cultural influence in product usage. International Journal of Human-Computer Interaction, 12(3&4), 327-345.

Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies, 64(2), 79 - 102.

Hornbaek, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. Interacting with Computers 20, 505–514.

Hornbæk, K. (2009). Dogmas in the assessment of usability evaluation methods. Behaviour & Information Technology, 1-15.

Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. Paper presented at the CHI, Portland, Oregon, USA.

Horton, W. (2005). Graphics: The not quite universal language. In N. Aykin (Ed.), Usabillity and Internationalization of Information Technology (pp. 157-188): Lawrence Erlbaum.

Hvannberg, E. T., Law, E. L.-C., & Lárusdóttir, M. K. (2007). Heuristic evaluation: Comparing ways of finding and reporting usability problems. Interacting with Computers, 19(2), 225-240.

Haak, M. J. v. d., Jong, M. D. T. d., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behaviour & Information Technology, 22(5), 339–351.

Iivari, N. (2004). Enculturation of user involvement in software development organizations - An interpretive case study in the product development context. Paper presented at the NordiCHI Tampere, Finland.

Iivari, N. (2006). Discourses on "culture" and "usability work" in software product development. University of Oulu.

Iivari, N., & Abrahamsson, P. (2002). The interaction between organizational subcultures and user-centered design - A case study of an implementation effort. Paper presented at the Proceedings of the 35th Hawaii International Conference on System Sciences.

Ito, M., & Nakakoji, K. (1996). Impact of culture on user interface design. In E. M. d. Galdo & J. Nielsen (Eds.), International User Interfaces (pp. 105-126). New York, Chichester, Brisbane, Toronto, Singapore: John Wiley & Sons, Inc.

Jacobsen, N. E. (1999). Usability evaluation methods: The reliability and usage of cognitive walkthrough and usability test. Unpublished ph.d., University of Copenhagen, Copenhagen.

Jansen, R. G., Wiertz, L. F., Meyer, E. S., & Noldus, L. P. J. (2003). Reliability analysis of observational data: Problems, solutions, and software implementation. Behavior Research Methods, Instruments, & Computers, 35(3), 391-399.

Ji, L.-J., Peng, K., & Nisbett, R. E. (2000). Culture, control, and perception of relationships in the environment. Journal of Personality and Social Psychology 78(5), 943-955.

Ji, L.-J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. Journal of Personality and Social Psychology, 87(1), 57-65.

Jokela, T. (2004). Evaluating the user-centredness of development organisations: conclusions and implications from empirical usability capability maturity assessments. Interacting with Computers, 16(6), 1095-1132.

Kaminsky, S. K. (1992). Test early, test often-A formative usability KH for writers. Paper presented at the SIGDOC

Karsh, B. T. (2004). Beyond usability: designing effective technology implementation systems to promote patient safety. British Medical Journal, 13(5), 388.

Keenan, S. L., Hartson, H. R., Kafura, D. G., & Schulman, R. S. (1999). The usability problem taxonomy: A framework for classification and analysis. Empirical Software Engineering, 4(1), 71-104.

Kim, H. S. (2002). We talk, therefore we think? A cultural analysis of the effect of talking on thinking. journal of Personality and Social Psychology, 83(4), 828-842.

Kim, Y. Y., & Gudykunst, W. B. (1988). Theories in Intercultural Communication. Newburry Park: Sage Publications.

Kimmel, P. R. (2000). Culture and conflict. The handbook of conflict resolution: Theory and practice, 453-474.

Kittler, M. G. (2009). How cultural context interferes with communication: A synthesis of Hall's concept of High- vs. Low-Context-Cultures and Krippendorff's information theory. Retrieved on 2009-2-10 from http://www.allacademic.com/meta/p_mla_apa_research_citation/0/9/2/3/3/p92335_index.html.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. International Journal of Human-Computer Studies, 60(5-6), 599-620.

Koch, S. C., & Zumbach, J. (2002). The use of video analysis software in behavior observation research: Interaction patterns in task-oriented small groups. Paper presented at the Forum: Qualitative Social Research (ISSN 1438-5627).

Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. Psychological Bulletin, 133(3), 464-481.

Krahmer, E., & Ummelen, N. (2004a). Thinking about thinking aloud-A comparison of two verbal protocols for usability testing. IEEE Transactions on Professional Communication 47(2), 105-117.

Krahmer, E., & Ummelen, N. (2004b). Thinking about thinking aloud-A comparison of two verbal protocols for usability testing. http://fdlwww.uvt.nl/~krahmer/Pubs%5Cieee2.pdf, 1-17.

Krippendorff, K. (1986). Informatoin theory: structural models for qualitative data. California, London, New Delhi: SAGE Publications, Inc. .

Krippendorff, K. (2004). Content analysis: An introduction to its methodology (Second Edition ed.). Thousand Oaks, London, New Delhi: Sage Publications: International Educational and Professional Publisher.

Kundel, H. L., & Polansky, M. (2003). Measurement of Observer agreement. Radiology. Retrieved January 21, 2009 from http://radiology.rsnajnls.org/cgi/reprint/228/2/303, 303-308.

Kuniavsky, M. (2003). Observing the user experience- A practitioner's guide to user research. San Francisco: Morgan Kaufmann Publishers, An Imprint of Elsevier.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. . Biometrics, Cited in http://www.musc.edu/dc/icrebm/kappa.html Retrieved on January 21, 2009, 33, 159-174.

Lavelli, M., & Poli, M. (1998). Early mother-infant interaction during breast-and bottle-feeding. Infant behavior and Development, 21(4), 667-683.

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. Behaviour & Information Technology, 16(4), 246-266.

Law, E. L.-C., & Hvanneberg, E. T. (2004). Analysis of combinatorial user effects in international usability tests. Paper presented at the CHI 2004, Vienna, Austria

Law, L.-C., & Hvanneberg, E. T. (2002). Complementarity and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. Paper presented at the NordiCHI 2002, Aarhus, Denmark.

Lee, J.-J., & Lee, K.-P. (2007). Cultural differences and design methods for user experience research: Dutch and Korean participants compared. Paper presented at the Proceedings of the 2007 conference on Designing pleasurable products and interfaces Helsinki, Finland

Lehman, D. R., Chiu, C.-y., & Schaller, M. (2004). Psychology and culture. Annu. Rev. Psychol., 55, 689–714.

Leventhal, L. M., & Barnes, J. A. (2007). Usability Engineering: process, products, and examples. Upper Saddle River, New Jersey: Pearson Education, Inc.

Levy, J., Wubbels, T., Brekelmans, M., & Morganfield, B. (1997). Language and cultural factors in students' perceptions of teacher communication style. International Journal of Intercultural Relations, 21(1), 29-56.

Lidwell, W., Holden, K., & Butler, J. (2003). Universal principles of design. Massachusetts: Rockport Publishers, Inc.

Lindgaard, G. (2009). Early traces of usability as a science and as a profession. Interacting with Computers.

Littlemore, J. (2001). An empirical study of the relationship between cognitive style and the use of communication strategy. Applied Linguistics, 22(2), 241-265.

Lucy, J. A. (1992a). Grammatical categories and cognition: A case study of the linguistic relativity hypothesis: Cambridge University Press.

Lucy, J. A. (1992b). Language diversity and thought: A reformulation of the linguistic relativity hypothesis: Cambridge University Press.

Lull, J. (2001). Culture in the communication age. London and New York: Routledge.

Luzio, D. A., Gunthner, S., & Orletti, F. (2001). Culture in communication--Analyses of intercultural situations. Amsterdam/ Philadelphia: John Benjamins publishing company.

Maloney, L. T. (2003). Statistical decision theory and evolution. Trends in Cognitive Sciences 7(11), 473-475.

Marcus, A. (2005). User interface design and culture. In N. Aykin (Ed.), Usabillity and Internationalization of Information Technology (pp. 51-78).

Marcus, A., & Gould, E. W. (2000). Crosscurrents: cultural dimensions and global Web user-interface design. interactions 7(4), 32 - 46

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data, A model comparison perspective (Second Edition ed.). Mahwah, New Jersey Lawrence Erlbaum Associates.

Mayhew, D. J., & Bias, R. G. (2005). Cost-justifying usability engineering for cross-cultural user interface design. In N. Aykin (Ed.), Usabillity and Internationalization of Information Technology (pp. 213-253): Lawrence Erlbaum.

Michael, T. (2009). One-way Analysis of variance in SPSS. http://aueb.academia.edu/documents/0008/6409/One-Way_Analysis_of_Variance_in_SPSS.pdf retrieved on 2-2-2009.

Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. Behaviour & Information Technology, 23(1), 65–74.

Mullany, M. J., Tan, F. B., & Gallupe, R. B. (2007). The impact of analyst- User cognitive style differences on user satisfaction. Paper presented at the 11th Pacific- Asia Conference on Informatoin Systems.

Nawaz, A., & Clemmensen, T. (2007). Cultural differences in the structure of categories among users of clipart in Denmark and China. Paper presented at the The Seventh Danish HCI Research Symposium (DHRS 2007), IT University, Copenhagen, Denmark.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. Paper presented at the In Proceedings of the ACM CHI' 92 conference, New York.

Nielsen, J. (1993). Usability Engineering[M]: New Jersey: AP Professional.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. Paper presented at the INTERCHI'93, Amsterdam, The Netherlands.

Nisbett, R. E. (2003). The geography of thought. London: Nicholas Brealey Publishing.

Nisbett, R. E. (2004). Cognition and perception East and West. Paper presented at the 28th International Congress of Psychology in Beijing.

Nisbett, R. E., & Masuda, T. (2003). Cultural and point of view. PNAS 100(19), 11163-11170.

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: holistic versus analytic perception. Trends in Cognitive Sciences 9(10).

Nisbett, R. E., & Norenzayan, A. (2002a). Cultural and Cognition. In D. L. Medin (Ed.), Stevens' Handbook of Experimental Psychology, third edition.

Nisbett, R. E., & Norenzayan, A. (2002b). Culture and Cognition. In Chapter for D. L. Medin (Ed.). Stevens' Handbook of Experimental Psychology, Third Edition.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. Psychological Review 108(2), 291-310.

Norman, K. L., & Panizzi, E. (2006). Levels of automation and user participation in usability testing. Interacting with Computers, 18(2006), 246-264.

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice?: An explorative study of think-aloud testing Paper presented at the Proceedings of the 6th ACM conference on Designing Interactive systems, 209 - 218 USA.

Observer. (2005). The Observer XT.

Onibere, E. A., Morgan, S., Busang, E. M., & Mpoeleng, D. (2001). Human-computer interface design issues for a multi-cultural and multi-lingual English speaking country -- Botswana. Interacting with Computers, 13(4), 497.

Punch, K. F. (2005). Introduction to social research-quantitative and qualitative approaches. London: SAGE Publications.

Rau, P.-L. P., Choong, Y.-Y., & Salvendy, G. (2004). A cross cultural study on knowledge representation and structure in human computer interfaces. International Journal of Industrial Ergonomics (34), 117-129.

Riding, R. J., & Rayner, S. G. (2000). International perspectives on individual differences: volume 1: cognitive styles. Stamford: Ablex Publishing Corporation.

Rogers, E. M., Hart, W. B., & Milke, Y. (2002). Edward T. Hall and the history of intercultural communication: The United States and Japan. Keio Communication Review, 24, 3-26.

Rose, K., & Zuhlke, D. (2005). Localization of user-interface-design for mainland China: Empirical study results and their description in a localization model and language issues. In X. Ren & G. Dai (Eds.), Evolution of the human-computer interaction (pp. 87-101). New York: Nova Science Publishers.

Sacher, H., Tng, T.-H., & Loudon, G. (2001). Beyond translation: Approaches to interactive products for Chinese consumers. International Journal of Human-Computer Interaction, 13(1), 41.

Sacks, H. (1994). On doing being ordinary. Language, 413-429.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 696-735.

Sadler-Smith, E., & Riding, R. (1999). Cognitive style and instructional preferences. Instructional Science, 27(5), 355-371.

Samovar, L. A., & Porter, R. E. (2003). Communication between cultures. Nelson, Canada: Wadsworth, a division of Thomson Learning, Inc.

Sanchez-Burks, J., Lee, F., Nisbett, R. E., Choi, I., Zhao, S., & Koo, J. (2003). Conversing across cultures: East–West communication styles in work and nonwork contexts. Journal of Personality and Social Psychology, 85(2), 363-372.

Sanchez-Burks, J., Nisbett, R. E., & Ybarra, O. (2000). Cultural styles, relational schemas and prejudice against outgroups. University of Michigan.

Sereno, K. K., & Mortensen, C. D. (1970). Foundations of communication theory. New York, Evanston, and London: Harper & Row, Publishers.

Shen, S.-T., Woolley, M., & Prior, S. (2006). Towards culture-centred design. Interacting with Computers, 18(4), 820.

Shi, Q. (2008a). A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. Paper presented at the NordiCHI.

Shi, Q. (2008b). A study of usability problem finding in cross-cultural thinking aloud usability tests. Paper presented at the Cultural Usability and Human Work Interaction Design – techniques that connects: Proceedings from NordiCHI 2008 Workshop.,

Shi, Q., & Clemmensen, T. (2007). Relationship model in cultural usability testing. Paper presented at the HCI International, 2007, Beijing.

Shi, Q., & Clemmensen, T. (2008). Communication patterns and usability problem finding in cross-cultural thinking aloud usability testing. Paper presented at the CHI 2008, Florence.

Skov, M. B., & Stage, J. (2005). Supporting problem identification in usability evaluations. Paper presented at the Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, Canberra, Australia

Slobin, D. I. (1991). From "thought" and language" to "thinking for speaking". In J. J. Gumperz & S. C. Levinson (Eds.), Rethinking linguistic relativity (pp. 70-96): Cambridge University Press.

Smagorinsky, P. (1998). Thinking and speech and protocol analysis. Mind, Culture, and Activity, 5(3), 157-177.

Smagorinsky, P. (2003). Rethinking protocol analysis from a cultural perspective. Annual Review of Applied Linguistics, 21, 233-245.

Smith, A., Dunckley, L., French, T., Minocha, S., & Chang, Y. (2004). A process model for developing usable cross-cultural websites. Interacting with Computers, 16(1), 63.

Smith, A., & Yetim, F. (2004). Global human-computer systems: Cultural determinants of usability. Editorial. Interacting with Computers, 16, 1-5.

Spencer-Oatey, H. (2000). Culturally speaking. London and New York: Biddles Ltd, Guildford and King's Lynn.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. . Practical Assessment, Research & Evaluation, 9(4). Retrieved January 21, 2009 from http://PAREonline.net/getvn.asp?v=9&n=4 . .

Sun, H. (2003, 5-10 April). Cultural usability: A localization study of mobile text messaging use. Paper presented at the CHI2003, Florida.

Sun, H. (2004). Expanding the scope of localization: A cultural usability perspective on mobile text messaging use in American and Chinese contexts. Unpublished PhD, New York.

Sun, H. (2006). An article by the 2005 Award Winner CCCC outstanding dissertation Award in Technical Communication: The triumph of users: Achieving cultural usability goals with user localization. Technical Communication Quarterly, 15(4), 457.

Sun, X., & Shi, Q. (2007). Language issues in cross cultural usability testing: A pilot study in China. Paper presented at the HCI International 2007, Beijing.

Tamler, H. (2001). High-tech versus high-touch: The limits of automation in diagnostic usability testing. http://www.htamler.com/papers/techtouch/.

Tamler, H. (2003). High-tech versus high-touch: The limits of automation in diagnostic usability testing. User Experience, 2(4), 18-22.

Tullis, T., & Albert, B. (2008). Measuring the user experience- Collecting, analyzing, and presenting usability metrics. Burlington: Morgan Kaufmann Publishers, Elsevier Inc.

Vatrapu, R., & Pérez-Quiñones, M. A. (2006). Culture and usability evaluation: The effects of culture in structured interviews. Journal of Usability Studies, 2006, 1(4), 156-170.

Vatrapu, R. K. (2007). Technological intersubjectivity and appropriation of affordances in computer supported collaboration. University of Hawai'I.

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. Paper presented at the Conference on Human Factors in Computing Systems.

Vygotsky, L. S. (1962). Thought and Language. Cambridge: Massachusetts Institute of Technology.

Vöhringer-Kuhnt, T. (2002). The influence of culture on usability. Master thesis., Berlin, Germany.

Weber, R. P. (1990). Basic content analysis (second edition): Sage Publications.

Wilson, C. E. (2007). The problem with usability problems: Context is critical. Interactions, 14 (5), 46-47, 50.

Witkin, H. A. (1967). A cognitive-style approach to cross-cultural research. International Journal of Psychology, 2(4), 233-250.

Woolrych, A., & Cockton, G. Why and when five test users aren't enough. http://www.netraker.com/nrinfo/research/FiveUsers.pdf.

Yammiyavar, P., Clemmensen, T., & Kumar, J. (2008). Influence of cultural background on non-verbal communication in a usability testing situation. International Journal of Design, 2(2), 31-40.

Yeo, A. W. (1996). Cultural user interfaces: a silver lining in cultural diversity. SIGCHI Bull., 28(3), 4-7.

Yeo, A. W. (1998a). Cultural effects in usability assessment. Paper presented at the CHI 98 conference summary on Human factors in computing systems.

Yeo, A. W. (1998b). Cultural effects in usability assessment. Paper presented at the CHI 98, Doctoral Consortium.

Yeo, A. W. (2001a). Global-software development lifecycle: An exploratory study. Paper presented at the CHI 2001.

Yeo, A. W. (2001b). Global-software development lifecycle: an exploratory study. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Yoshioka, T., & Herman, G. (1999). Genre taxonomy: A knowledge repository of communicative actions. Paper provided by MIT Center for Coordination Science in its series Working Paper Series with number 209., 26.

Zahedi, F., Van Pelt, W. V., & Srite, M. (2006). Web documents' cultural masculinity and femininity. Journal of Management Information Systems, 23(1), 87.

8 Appendices

Appendix 1: Consent for the evaluators and users

Consent

The <HCI research group in Denmark and China> has developed a prototype of a ‘wedding invitation application’, which is combined with Microsoft Clip Organiser in Microsoft Word. Now we want you to use the prototype and give us comments of your feelings of the application and suggestions of the changes in order to make such application easy to use. The study takes place in Usability Lab of the HCI research Group. We will record information about how you use this application. We will also ask you to fill out questionnaires. All or some of your work with the site will be videotaped and recorded. We would like to use these videotaped and recorded sessions to make changes in the application and also for educational and research purposes. Please sign this consent form to allow us to use these materials for the above mentioned reasons. Thanks!

Location: <Copenhagen/ Beijing>
Signature: ___________________________
Name: _____________________________
Date: ________________

Appendix 2: Background questionnaire for the users

Background questionnaire for the user

Part 1: Demographic information

Age: ___________ years

Nationality: _______________

Country of residence: _______________

Were you raised in this country: __ yes, __ no

Years of education (elementary school + middle school + …): ___________ years

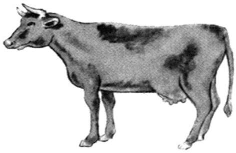
Please choose the option that corresponds to you for the following questions.

Q1. What is your first Language?
A. English; B. Danish; C. Chinese; D. Hindi; E. other, please specify______________
Q2. Which language is the official language in your office? Like making presentation, writing report to your boss…..
A. English; B. Danish; C. Chinese; D. Hindi; E. other, please specify______________
Q3. Which language do you prefer while working with Computers?
A. English; B. Danish; C. Chinese; D. Hindi; E. other, please specify______________
Q4. If ask you to communicate with people in English, you think you can speak English:
A. Very bad; B. bad; C. Normal; D. Good; E. Very good.
Q5. How often do you work with the software Microsoft Word?
A. Never; B. Rarely (few times in a month); C. Occasionally (once or twice in a week); D. Often (almost everyday); E. Very often (every day, or always use many times in one day).
Q6. While working with text in Microsoft Word which of the following options have you used? (You can choose more than one option.)
A. Font; B. Font Size; C. Font Colour; D. Highlighting; E. Style; F. Alignment; G. Bulleting and Numbering.
Q7. In Microsoft Word, have you used the function of inserting pictures?
A. Yes; B. No
Q8. While working with the Microsoft Word, have you tried using the picture toolbar?
A. Yes; B. No
Q9. While working with pictures and images in Microsoft Word, which of the following options have you used?
A. Brightness; B. Contrast; C. Rotate; D. Text Wrapping. E. None
Q10. Have you ever received a wedding invitation from your friends/ relatives?
A. Yes; B. No
Q11. Have you ever seen a Danish (tests in Denmark) wedding invitation? (Change Danish to Chinese for the tests in China)
A. Yes; B. No

Part 2

Please select only one answer (A, B or C) according to your feeling. There is no right or wrong answer. Please stick to your first impression and also briefly tell us the reason why you think so.

1) If you are asked to categorize the following three objects-- grass, chicken and cow, which two could be put together? (   )

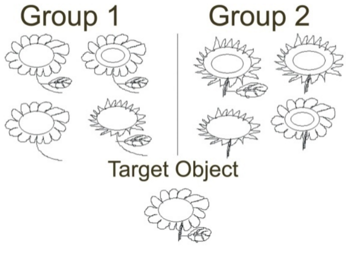


A. Chicken and cow
B. Grass and chicken
C. Cow and grass

The reason is:____.

2) Please indicate whether the target object is more similar to the objects of Group 1 or Group 2. (   )

The reason is :____.



A. The target object is more similar to Group 1
B. The target object is more similar to Group 2

Appendix 3: Background questionnaire for the evaluators

Background questionnaire for the evaluator

Part 1: Demographic information

Age: ___________ years

Nationality: _______________

Country of residence: _______________

Were you raised in this country: __ yes, __ no

Years of education (elementary school + middle school + …): ___________ years

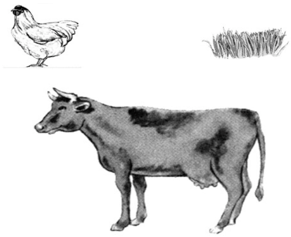
Please choose the option that corresponds to you for the following questions.
Q1. What is your first Language?
A. English; B. Danish; C. Chinese; D. Hindi; E. other, please specify______________
Q2. Which language is the official language in your office? Like making presentation, writing report to your boss…..
A. English; B. Danish; C. Chinese; D. Hindi; E. other, please specify______________
Q3. Which language do you prefer while working with Computers?
A. English; B. Danish; C. Chinese; D. Hindi; E. other, please specify______________
Q4. If ask you to communicate with people in English, you think you can speak English:
A. Very bad; B. bad; C. Normal; D. Good; E. Very good.
Q5. How often do you work with the software Microsoft Word?
A. Never; B. Rarely (few times in a month); C. Occasionally (once or twice in a week); D. Often (almost everyday); E. Very often (every day, or always use many times in one day).
Q6. While working with text in Microsoft Word, which of the following options have you used? (You can choose more than one option)
A. Font; B. Font Size; C. Font Colour; D. Highlighting; E. Style; F. Alignment; G. Bulleting and Numbering.
Q7. In Microsoft Word, have you used the function of inserting pictures?
A. Yes; B. No
Q8. While working with the Microsoft Word, have you tried using the picture toolbar?
A. Yes; B. No
Q9. While working with pictures and images in Microsoft Word, which of the following options have you used?
A. Brightness; B. Contrast; C. Rotate; D. Text Wrapping. E. None
Q10. How many Usability Test you have conducted before as an evaluator?
A. None; B. 1-2; C. 3-5; D. more than 5
Q11. How many years have you been working as a HCI Specialist/ Human Factors Engineer/ Usability?
A. < 1 year; B. 1-2 years; C. 2-4 years; D. more than 4 years.
Q12. Have you ever received a wedding invitation from your friends/ relatives?
A. Yes; B. No
Q13. Have you ever seen a Danish (tests in Denmark) wedding invitation? (Change Danish to Chinese for the tests in China)

A. Yes; B. No

Part 2

Please select only one answer (A, B or C) according to your feeling. There is no right or wrong answer. Please stick to your first impression and also briefly tell us the reason why you think so.

1) If you are asked to categorize the following three objects-- grass, chicken and cow, which two could be put together? ( )

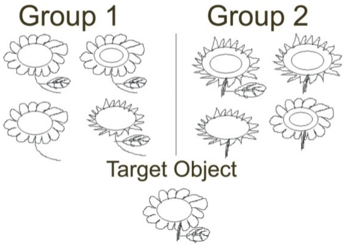


A. Chicken and cow
B. Grass and chicken
C. Cow and grass

The reason is:____.

2) Please indicate whether the target object is more similar to the objects of Group 1 or Group 2. ( )

The reason is :____.



A. The target object is more similar to Group 1
B. The target object is more similar to Group 2

Appendix 4: Usability problem list for the users

Usability Problem List For Danish Users in the test (Change "Danish" to "Chinese" in the tests in China)

Please mark the problems which you encountered during the test, which means you just mark the problems that you mentioned or happened during the test. Do not need to mark the problems that you now think it is a problem, but did not encounter when you were doing the test. Please also rate the severity of the usability problems that you choose: **M** for minor problems, **I** for important problems, **C** for critical problems.

N	Potential problems	Encountered problems	Severity (minor, important, critical)
1	Some images are neither used for wedding, nor Danish culture		
2	Some images are not used in Danish wedding invitation, which means images are not related to wedding.		
3	Images represent other culture which will never used for Danish people in their wedding		
4	Inappropriate invitation text, which means the content/ expression of the invitation is not appropriate for Danish wedding invitation		
5	Task support problem, which is related to the task, such as not so clear of the task, inappropriate tasks, or not includes better tasks.		
6	Image moving problem (in Word), which is the problem related to move images on Word, such as hard to move images or cannot put it on the suitable place.		
7	Locate Clip Art subfolder problem, which is problem related to find/ open/ notice the folder of images, texts or backgrounds		
8	Problem of the Clip Art Search collection, which is the one on the right sides of MS Word when you click "insert picture-Clip Art".		
9	Image collection range problem (too few), such as the range of images in the image folder are not enough to select from		
10	Image preview problems in the folder, such as do not know how to preview the images, preview size too small, cannot enlarge preview size……		
11	Image layout problem (in Word), such as problem with text wrapping tool, hard to make the text over image….		
12	Image formatting problem, such as editing size, contrast, brightness, color and so on.		
13	Image thumbnail size is small		
14	Clipart match problem: the background and images are not matched.		
15	Background creating problem, such as do not know how to make a background, cannot make a satisfied background, or do not know how to make a picture as a background.		
16	Application concept comprehension problem, such as unable		

	to comprehend meaning and purpose of the folders that are provided		
17	Problem with inserting image from the Clip Art to Word, such as not convenient to do that		
18	Thumbnail image resolution problem, such as too low resolution, do not know the resolution.		
19	Problem of no category of images: the images should be categorized into some sub-categories.		
20	Keywords to find images are not clear		
21	Image collection aesthetic problem: images are not pretty/ romantic/ good for wedding invitation		
22	Image selection problems of too many images		
23	Invitation text quality problem: the text may be required more consideration to make it better.		
24	It is not easy to choose the appropriate font(s) for the text, such as text font, the font style, and the size of the words		
25	Not easy to do the format and layout for the text, such as the space between the words, the place of the text on the paper……		
26	Inappropriate invitation background, such as some backgrounds are not for Danish wedding invitation (not for the culture, or not for wedding)		
27	Background quality problems: backgrounds are not pretty/ good enough		
28	Background editing problem, such as it took some time to edit the color, brightness, contrast and so on.		
29	Problem with print preview: such as cannot see the background from the print preview		
30	Background collection range problem (too few): such as the backgrounds are not enough		
31	Background collection range problem (too many)		
32	No background category		
33	Background preview problem: cannot preview the backgrounds, should open files one by one		
34	Problem of making the text vertical: wanted to make the text vertical, but it was not easy.		
35	General word processing problems, which are other problems related to MS word not related to the manipulation of the Collection, such as hard to find the picture toolbar, not happy with some functions in Word......		
36	Others (please specify):		

Appendix 5: Usability problem form for the evaluators

Usability Problem form for the evaluator

Please write down the usability problems that you found in this test session. Please also rate the severity of the problems: **M** for minor problems, **I** for important problems, **C** for critical problems.

Tasks	Usability Problems	Severity (M, I, C)
Backgrounds choosing (includes user's evaluation/ opinion of backgrounds)	1. 2. 3. 4. …..	
Background editing (like the colour, brightness, contrast and so on)	1. 2. 3. 4. …..	
The invitation texts that were provided.	1. 2. 3. 4. …..	
Text font editing (such as the text font, the font style, and the size of the words/ characters)	1. 2. 3. 4. …..	
Text colour editing	1. 2. 3.	

	4. …..	
Text formatting and layout (such as the space between the words, the emphasis mark, the place of the text on the paper and so on)	1. 2. 3. 4. …..	
Images that were provided (evaluation/ opinion from the user, such as cultural issues, wedding issues, aesthetic issues, and so on)	1. 2. 3. 4. …..	
Image choosing (includes preview, choose the image to Word, and so on)	1. 2. 3. 4. …..	
Image editing (such as the size, brightness, contrast, layout, moving, and so on)	1. 2. 3. 4. …..	
Others (such as problems with MS Word)	1. 2. 3. 4. …..	

Appendix 6: Short interview with the users after the test

Short Interview with the user:

1. How do you think using English to think aloud? Did you express clear what you wanted to say in English?

2. If using a foreign evaluator (or since just now the evaluator was a foreign person), which of the following statements fits your situation?
    4) I will explain more to the foreign evaluator about what I am doing, since I think he/she may not understand. I hope he/ she can understand why I am choosing this not that.
    5) I just do the task and speak out as required. I do not talk more to the foreign evaluator, since he/she cannot understand.
    6) Neither. I behave the same to local and foreign evaluators.

Appendix 7: Short interview with the evaluators after the test

Short interview with the evaluator after doing all the four tests:

1) What are the criteria of choosing usability problems in this study (reasons of deciding what are usability problems)?

2) What are the criteria of deciding the severity of the usability problems? How to decide which problems are minor problems, which are important problems and which are critical problems?

3) If you think doing tests in your own culture is different from doing tests in other cultures, please tell me what the difference is and how to avoid cultural bias from your own experience?

Appendix 8: An example of the email for recruiting Danish users

Welcome to Attend My Study (Danish people only)

Dear friends,

I am a PhD student in Department of Informatics, CBS. I will be very appreciated if you could attend my study. The study will take you around 1 hour. We will give you a nice gift as thanking you (now the gift is a box of 2 CDs with very nice traditional Chinese music, and you can also choose a bottle of wine).

What will you do in the study?

We are going to do a very interesting usability test. It is not testing you, but testing an application called "wedding invitation application". We want to see whether this application is good enough for you to make wedding invitations. **You do not need to prepare anything before coming here.** You just come and we will tell you how to do it. It is very easy for you. It includes three parts:

1. A short questionnaire about your background;
2. Making a wedding invitation you may want to use by using the application we provide;
3. A short interview after the test.

Requirement of the participants:

1. **Danish people** who are born and grow up with Danish parents in Denmark and between 22 to 40 years old.
2. The study will be in English, so hope you can **speak English fluently**.
3. **Use Microsoft Word often** and be familiar with it.

Time: In the following 3 months. We can discuss the time according to your schedule

Place of the study: Howitzvej 60, office 3.16, DK-2000 Frederiksberg

If you are interested in it: **qs.inf@cbs.dk**

Look forward to hearing from you!

Best regards,

Qingxin Shi
PhD student
Department of Informatics, CBS, Howitzvej 60, office 3.16, DK-2000 Frederiksberg
Mobile: (+45) 7242 1536

Appendix 9: The researcher's instruction to the evaluators

Instruction to the evaluator

1. Thanks for agreeing to participate in this experiment. The experiment is about thinking aloud usability testing.
2. The experimental session you are about to start consists of four parts. First, I will ask you to fill in a short background questionnaire. Second, you conduct the usability test with the user. Third, please write down the usability problems that you experienced in the test, and the severity of the problems. Fourth, you will be interviewed by the researcher about the criteria of choosing usability problems and the criteria of deciding which problems are minor problems, which are serious problems and which are critical problems.
3. Since you have the experience of being an evaluator, you know what thinking aloud usability testing is. I will not introduce it in detail now.

   Thinking aloud involves asking participants to verbalize their thoughts as they are performing a set of specified tasks. Users are asked to say whatever they are looking at, thinking, doing, and feeling, when they are doing their tasks.
4. Your task is to find as many usability problems as possible. You could communicate with the users during the test as you usually do. If the user asks questions about the task or needs help in doing the tasks, as long as you think the usability problems are not covered, you can give the user some information if you want to. After each test, I will ask you to write down the usability problems in each test and give severity of the problems. You need to write down every problem you find in each test, even though some of them you have written in the previous tests, if you find it again in this test, you need to write it down again.
5. The usability problems are not only function problems, such as not clear of how to do it, but also including the content of the application, such as the users are not satisfied with the wedding images, backgrounds or texts we provided.
6. In the test, first, you need to give instructions to the user, and then ask them to fill in a questionnaire (see the document of the questionnaires for users). Second, you need to introduce what the thinking aloud is. You show them an example of thinking aloud and then ask the user to do an exercise of thinking aloud (all this information you could find from the document of "evaluator's instruction to the user").
7. The task for the users is to make a wedding invitation they would like to use in their own wedding using the materials we provide. The wedding invitation prototype includes three parts: background, texts and images. There are 8 backgrounds and 1 blank background. If they are not satisfied with the 8 backgrounds provided, they could use the blank background to make their own background. We also provided 4 texts about wedding invitation. The user could choose one and make some changes since it is their own wedding. If they are not satisfied with the texts, they could write their own. We collect some images related to wedding and put them in the Clip Art organizer. No matter the users like the images or not, they have to pick at least one and put it on the invitation.
8. Actually the prototype is just a collection of backgrounds, texts and images using Microsoft Word. I will show you before the test. The Microsoft Word and invitation texts are in Danish/Chinese, but the users will think aloud in English.

9. Your participation in the experiment is anonymous, which means your name will not be recorded.

Appendix 10: The evaluator's instruction to the user

Evaluator's instruction to the user

A company which helps people to design wedding invitations wants to combine the function of making wedding invitation to Microsoft Word. They made a prototype called "wedding invitation clipart" which includes three parts, images folder, backgrounds folder and texts. The images folder is combined with Microsoft Clip Art Organiser. In that window you will find on the left hand side there are many folders (My collections, Office Collection and Web Collection). They have incorporated their "Wedding invitation Clip Art" in this clip Organiser by adding a sub folder in My Collections. You could see some sub-folders named "Chinese wedding invitation images", "Danish wedding invitation images". From these folders, you can choose images. Backgrounds and texts are also provided in different folders which could be used to make the wedding invitation.

Since the company wants to see whether the folders are good enough for people to make wedding invitations, they ask us to do the usability tests. That's why we invite you here to use them. Since we want to see whether the Danish (Chinese in the tests in China) materials are good for you, please just use the Danish (Chinese in the tests in China) wedding invitation images, backgrounds, and texts folders, which I have already opened for you.

The whole usability testing includes three parts, first, a short questionnaire, second, usability test and third questionnaire and interview after the test given by the researcher (the evaluator does the first two parts, and the researcher does the third part).

Your participation in the experiment is anonymous. That is, your name will not be recorded.

First please fill out the Consent (see the Consent).

Thank you. Now please fill out the short questionnaire (see the questionnaire for users).

When finishing the consent and questionnaire, the evaluator will introduce what thinking aloud is, show the user an example of thinking aloud and ask the user to do an exercise.

Instruction of the Thinking Aloud Usability Test to the User

Thinking Aloud is the method which asks you to say whatever you are looking at, thinking, doing, and feeling, when you are doing your task. In a simple word, it means speaking out everything in your mind when you are doing the task. For example, if ask you to open Microsoft PowerPoint, you are opening MS PowerPoint and at the same time saying "I go to start and find Microsoft, and here is PowerPoint, I click it, and now it is open."

Please do an exercise: "open Windows Media player and choose a song to listen"

Ok, you did very well. Let's start our test.

Introducing the application

This is a Danish/Chinese wedding invitation Clipart Organizer. You can choose images/ pictures/ symbols from here. This is the background folder, you can choose a background/ paper for the wedding invitation if you like. There are some wedding invitation text samples. You can choose any sentence to make the wedding invitation. Of course, if you are not satisfied with the background or text, you could make your own. (If they are not satisfied with the backgrounds or texts we provide, they could make their own. The images in the Clip Art folder should be chosen at least one)

About the task

Do not show the task list. There are some small pieces of paper with each task on. You give the tasks to the user one by one. And ask them to do each task while thinking aloud.

Appendix 11: Task list in the test

Task List

Imagine you will get married and you want to make a wedding invitation by yourself. Please make a wedding invitation that you want to use in your wedding by using the wedding invitation application we provide.

The general task is divided into eight sub-tasks:

1. You can choose a paper (background) for the wedding invitation if you want to.
2. You can edit the background you chose if you think it is necessary.
3. Please write the text that you want to appear on the invitation. (You can pick one from the text examples file, and you can also make some changes if you want to.)
4. Please choose the appropriate font(s) for the text (such as the font, the font style, the size and so on).
5. Please choose the color(s) for the text.
6. You are free to choose any kind of formatting and layout for the text and make it prettier.
7. Now choose one or some images from the Danish/Chinese wedding invitation Clipart sub-folder to make the wedding invitation look happy, colorful, and joyful, as this is for a wedding.
8. Please edit the images if you want to (such as the size, brightness, contrast, layout, and so on).

Appendix 12: Protocol for the tests

Protocol for the tests

There are some probing questions, besides these questions, you can also probe other questions during the test if you think it is necessary for you to get relevant usability problems. Please do the test as you usually do.

Task 1. You can choose a paper (background) for the wedding invitation if you want to.

Probing questions:

a). How do you think of the backgrounds that we provide?

b). Do you think the backgrounds are necessary for you to make wedding invitation?

c). Do you have any suggestions for the backgrounds folder?

d) Any other questions that need to be asked when you are observing the user choosing the backgrounds.

Task 2. You can edit the background you chose if you think it is necessary (like the colour, brightness, contrast and so on).

Probing questions:
a) Do you think it is easy to edit the colour/ brightness/ contrast…?
b) Any other questions that need to be asked when you are observing the user editing the backgrounds.
Task 3. Please write the text that you want to appear on the invitation. (You can pick one from the text examples file, and you can also make some changes if you want to.)
Probing questions:
a). How do you think of the texts that we provide?
b). Do you think the texts are necessary for you to make wedding invitation?
c). Do you have any suggestions for the texts document?
d) Any other questions that need to be asked when you are observing the user choosing and writing the text.
Task 4. Please choose the appropriate font(s) for the text (such as the font, the font style, the size and so on)
Probing questions:
Ask any questions that you think may be helpful in finding usability problems while you are observing the user choosing the text font, the font style, and the size of the words.
Task 5. Please choose the color(s) for the text.
You can probe any questions that you think may be helpful in finding usability problems.
Task 6. You are free to choose any kind of formatting and layout for the text and make it prettier. (such as the space between the words, the emphasis mark if necessary, the place of the text on the paper and so on, as long as you think the text has already pretty.)
You can probe any questions that you think may be helpful in finding usability problems.
Task 7. Now choose one or some images from the Danish/Chinese wedding invitation Clipart sub-folder to make the wedding invitation look happy, colorful, and joyful, as this is for a wedding.
Probing questions:
a). How do you think of the images that we provide?
b) Do you think the images are necessary for you to make wedding invitation?
c). Do you have any suggestions for the images folder?
d) Any other questions that need to be asked when you are observing the user choosing and inputting the images.
Task 8. Please edit the images if you want to (such as the size, brightness, contrast, layout, and so on).
You can probe any questions that you think may be helpful in finding usability problems when the user was editing the size, brightness, contrast, layout and so on.

Appendix 13: Transcribing instruction

Before doing the coding with Observer, the videos need to be transcribed into Excel. This is the instruction of transcribing the videos:

1) Transcribe the selected videos in Excel
2) Write down the time, the person and the speech;
3) Need to write down every word the user and evaluator said, including ok, sure, en…
4) If some words are not so clear, you just need to write down a word with the similar meaning you thought. If you don't know which word could be written, then from the video, maybe you could guess whether the user's meaning is negative or positive towards that application on that moment. For example, if the user said "it is too tacky, I will not use it". If you do not know the word tacky, but from the following sentence "I will not use it", you could guess it is negative description, then you just write "it is too "xx" (maybe negative thing), I will not use it". If you could not guess anything from the speech and can not get it, just leave it there and mark the word you did not get as "xx" with red colour.
5) Normally write down the person's speech in one time session. But sometimes you could write it down in different time sessions if there are short breaks or expressing different things: for example:

2:25	U	en, maybe I would use a photograph of you know, the actual couple getting married
2:35	E	the actual couple, yea, sure
2:38	E	but if you have to choose one or more

At time 2:35, the evaluator expressed "affirmative", at time 2:38, the evaluator expressed "instruction"

6) when the user or evaluator is talking, the other person is laughing or saying ok during the user/ evaluator talking, do it like this:

2:44	U	it is I guess, a little bit pink and tacky, but I would choose this one
	E	Haw-haw (laughing when the user is talking)
2:5	E	Yea

7) If the user is talking in a long period of time, in between the evaluator said "ok", we need to write down these short words. But some "ok" said by the evaluator frequently in a short time period and the user did not stop talking, then just write down the user's speech and write down how many "ok" the evaluator said during this talk. For example, the user said "I like everything with hearts on, but not too colourful, like this one, I also don't like pictures with wedding couples, I would use my own pictures". When the user was saying that, the evaluator said 3 "en" to agree, since the user did not stop when he/she was saying that, the time recorded may be from 10:02 to 10:15. In this session, for the evaluator, you write down the last "en" and mark"(2 more ok when the user was talking)"
8) In order to see the time clear, it is better to write down the time when the sentence starts. How to do that?

For short reply, such as “yea”, you could record the time when you heard it.
For short but longer than “yea”, such as “ok, I see”, you click stop, and write down the time as “the one you saw to deduct 1 or 2 seconds”, you need to go back to see the video and get the experience to see how much you need to deduct at the beginning and then do it in the following videos.
For longer sentence, you need to write down the sentence when it starts

9) When nobody was talking for more than 10 seconds, then record it as silence. You need to record it when the silence starts. For example:

10:20 U: “I like everything with hearts on, but not too colourful, like this one, I also don’t like pictures with wedding couples, I would use my own pictures. But in this case, I need to find a picture I like……”
10:34 silence
10:45 E: “could you tell me what you are thinking?”
It means the silence lasts for 11 seconds.

Appendix 14: Coding instructions

This is the coding instruction on how to do the coding in Observer. It includes two parts: the first part is “coding system and brief description”, the second part is the “explanation of the codes”. Please read it carefully. If you have any question during the coding, please write it down and discuss with me.

Part 1: Coding system and brief description

UP state event:

1. UPU: the problem is started by the user’s current behaviour or speech. The usability problem is from the current observation of the user’s behavior and listening to the user’s speech, such as “negative comments”, thinking aloud or user’s problem of doing the task.
2. UPE: the problem is started by the evaluator’s directed probing behaviours, without any negative comments from the user or without any problems of the user’s task performing. For example, the evaluator asked “how about the photos with other people?”

Each of them with 2 modifiers:

1) UP number: each problem related to the tasks of images is given a number.
   a. UP 1
   b. UP 2
   c. UP 3
   d. UP 4
   e. UP 5
   f. UP 6
   g. UP 7
   h. UP 8
   i. UP 9
   j. UP 10
   k. Two UPs are discussed/ talked at the same time and hard to distinguish
   l. UP finish (since it is state event, you need to have an end. It is defined as “default state event”, which means choosing another state in this category will stop the previous state in this category. For example, if you choose “UP 1” start, when you choose “UP 2”, you do not need to end “UP 1”. UP1 will be ended automatically when you choose UP 2. But sometimes UP 1 ended, but no other usability problem started, at that time you need to choose problem finish in order to end UP 1)

   Note. UP is for usability problem.

2) Usability problem severity: also from the usability problem list written by the evaluator.
   a. M (minor)
   b. I (important)
   c. C (critical)
   d. O (not a problem, with usability problem finish)

User’s state events: only one code with 3 modifiers

User’s state behaviours----the modifiers are:

1) Classic thinking aloud (Ericsson and Simon’s level 1 and 2 data, explained below)

2) Talk other than classic thinking aloud, such as comments, explanation, opinions by answering questions and so on

3) Silence (the duration of silence should be more than 10 seconds)

Point Event

1. **Users' point events behaviours:**
   1) Asking questions related to the task
   2) Short reply, such as ok, yea, no.
   3) Negative comments (take note of all the negative comments), such as the one may imply a usability problem.
   4) Positive comments (take notes), such as the users said they liked it or it was easy to do.
   5) Suggestions which imply how to improve the application or the user's wishes of the images or the application (take note of all the suggestions).
   6) Culture related comments: if the user mentioned in Denmark/ China…, or religions and any cultural issues, then code it as culture related comments.
   7) Others: any behaviour that may play a role in finding usability problems or potential problems, but not belonging to the above categories. It also includes the neutral comments, such as "it is ok".

   Notice:

   A. If a sentence includes both "culture related comment" and "suggestion"/"positive"/"negative", then code it both. For example, you could code the first half sentence as "culture related comment", and code the second half sentence as "suggestion" or "negative comment".

   B. For comments and suggestions, even though they were broken by the evaluator's "affirmative response" or "digging dipper probing," they were coded only once if they were mentioned at the same time and no other topic was in between. If the user's talk was broken by the evaluator's affirmative sentence, then just the first sentence was coded as "negative," "positive" or "suggestion."

   C. If the user's talk is broken by the evaluator's verbal behaviour, then you need to see whether the user has said the meaning to decide whether to code the first half sentence as "negative", ”positive" or ”suggestion". If the user has not said the meaning in the half sentence, and interrupted by the evaluator's talk, then code the first half sentence as "others". And code the following sentence after the evaluator's talk as "negative", ”positive", ”suggestion" or "short reply" according to the user's talk. For example,

U: "but a lot of it is a little bit, en…";	code it as "others"
E:”a bit too much";	code it as "clarify" in "probing"
U:”a bit too pink, and a bit too en…";	code it as negative comment
E:”a bit too tacky";	code it as "clarify" in "probing"
U:”yea".	code it as "short reply"

2. **Evaluator's point events behaviours:**
   1) Affirmative response, the short sentence/ words to act as an active listener, such as ok, yes, en, sure...
   2) Probing behaviour----modifier (a, b, c, d): all need to take notes

a. Directed probing (evaluator controlled probing). Here are the explanations:

   i. Basically, all the probing that are not followed with the user's current talk or behaviour

   ii. When the user did not notice something, the evaluator asked a specific part of the application and turned her/his attention on that part, such as "how do you think the images with other people's photo on it?"

   iii. Any questions with emotion if it was not followed with the user's current emotion, such as "you are not happy with the other people's photos?" Before that, there was no clear sign that the user was unhappy of that photo.

b. Act as reminders to think aloud or talk, such as "and now…?," "what are you thinking," "what are you looking for.", "what are you trying to do".

c. Digging deeper probing (user controlled probing), including asking for clarifying or speaking out the user's meaning by guessing:

   i. Basically, all probes followed with the user's current talk, behaviour or emotion

   ii. Ask the user to say it again or to clarify their meaning, such as "sorry, could you say it again?", "you said…., right?";

   iii. Talking the evaluator's understanding of the user's behaviour, including repeating the user's speech, in order to see whether he/ she understood the user.

   iv. The questions with the same emotion as the user's related emotion, such as "you are not so happy?"

   v. From the observation, the user had some problems of doing something and he/she talked a little, such as "oh", "my god", but not clearly. The evaluator asked the user whether he/she had the problem or how they thought of doing it

   vi. Digging deeper: such as the user said "I do not like these images", the evaluator asked "why?"

d. Others: including all the general questions and planned probes (Dumas & Loring, 2008) that are provided in the testing protocol. Besides, it also includes the questions that get the user's attention back to the task or emphasize the task, for example: are there any pictures you would use? The three questions provided in the testing protocol are: 1) How do you think of the images that we provide? 2) Do you think the images are necessary for you to make wedding invitation? 3) Do you have any suggestions for the images folder?

3) Classic reminders (take notes), such as please keep talking, please tell me what you are thinking.

4) Help behaviour: (take notes): One help is just coded once, even though there are many communications. The modifiers are:

   a. Try to help the user to figure out the problem--- do the help or willing to do the help

   b. No clear answer or help: do not give the user clear help or answer, but encourage the user to figure it out

5) Others: some important things you want to show other people (take notes), and it also includes the following behaviour, such as help behaviour lasts for a while, you code it as "help behaviour" at the beginning, and then code it as others if it continues in the following. Besides, when it is not clear to show the users' or evaluators' meaning, then also code it as "others".

Part 2: Explanation of some important codes

1. *When to code UPU and when to code UPE?*

   a. UPU: In these conditions, code the usability problem as UPU:

      i. From user's speech, such as "negative comments", thinking aloud, suggestions

      ii. From user's behaviour: it is from the user's task performing behaviour. When the user is doing a task, even he/she did not complain, but if the evaluator write down the problem from the observation of the user's difficulty of doing the task, then it is still coded as the UPU, since it is from the evaluator's observation. For example, the user is trying to insert an image, and spent some time on doing it without complaining anything. The evaluator observed and wrote the problem as "not obvious how to insert image in invitation"

      iii. If the problem is first from the user's behaviour, but the user did not complain, and the evaluator ask, it also belongs to this one. Because it is also from the user's behaviour, not the evaluator's existing thoughts. Even the communication is started from the evaluator's speech, but since it is easily to see the reason of asking the question is because of the user's behaviour. So it is still UPU.

      iv. If it is not directed probing, such as a general question "how do you think the images we provide", if the user said any negative things, then code it as UPU.

   b. UPE: The focus of this research is on the evaluator's problem finding behaviour. The problem may have been known by the evaluator from previous tests or their own mind, then the evaluator will use directed questions to see whether the user takes it is a problem or not. In these conditions, code the problem as UPE:

      i. The problem is started by the evaluator's directed probing without any negative comments by the user and without any problem that is doing by the user. For example, the evaluator asked "how about the photos with other people?"
      Directed questions are those pointing to a specific part of the application and it is not related to the user's current behaviour or comments.

      ii. If a problem has been found in this test before and the discussion of the

usability problem has been stopped, after a while, the problem is mentioned again by the evaluator without any clue of the user's current behaviour or speech, then code it as UPE.
A problem was found by the user first (UPU) and had been discussed by the user and evaluator, this problem period was over. After a while, the evaluator mentioned it again. This time, code it as UPE. So for one problem, it could be both UPU and UPE.
For example: the user said "I would not choose other people get married", and the evaluator said "ok". The problem of "photographs should not be of strangers" would be marked as "UPU" this time. Then they talked some other things and found some other problems. When the user was thinking and doing some other things, the evaluator said "the photographs you will definitely not use, photographs of other people", the use said "no, I wouldn't use". Here you need to mark it as "UPE", since it is not the user who mentions it this time. The problem has been in the evaluator's mind and the evaluator uses directed probing to mention it again, so it is coded as UPE.

2. *How to start and end a usability problem?*
    1) Start the usability problem just before the problem period:

Since usually the problem is together with evaluator or user's speech, in one time session, it is hard to code it as both problem and the behaviour, such as the evaluator said "the picture of other people's photograph you will definitely not use", in this situation, you will code this sentence as both the state event of "problem begin" and point event of "evaluator's directed probing behaviour". Since at one time point, you can only do one coding. The problem is state event, you need to know when it starts, so you just go back a little bit ahead of this sentence, and choose problem first, and then continue to code the following behaviour.
Usability problem should start before the whole situation, such as before the evaluator's general questions. For example, the evaluator asked "how do you think of the images", the user said "I do not like it, it is not personal". Then the UPU should start before the evaluator's question and this question is coded as "others" in "probing behaviour".

    2) How to end a usability problem:

The end of the usability problem: end it when this problem period is finished and before they begin another topic.

3. *How to code the two similar usability problems written down by the evaluator and discussed almost in the same communication session*
    1) If there are two similar problems emphasizing different things in the list written down by the evaluator, they happened/ discussed at almost the same time, how to mark it?

For example:
The two usability problems are: 1. pictures need transparent backgrounds; 2. not enough options for image manipulation in Word-e.g. removing backgrounds.

In this situation, actually, even though the two problems are similar and they happened at one communication session, you could still see which part of the communication emphasizes which problem. In this example, UP 1 is talked by the user as: "en, it is better if there is no background around the rose". UP 2 is talked in the following as how to do it. "maybe I would actually…. I don't know if that was possible. En…. This is background; I would like just this rose, without the square around it".

2) If it is very difficult to distinguish where is UP1 and where is UP2, then code it as "Two UPs are discussed/ talked at the same time and hard to distinguish", and take note of the two problems in the comments part. But in this situation, the two problems should have the same severity rank. If the two problems have different severity ratings, it is better not to choose the "Two UPs are discussed/ talked at the same time", but code them separately.

4. *How to code the user's state event?*

For state event, it is very important to know when it starts and when it ends

Start and end:

The start is coded just right before the event happens and the end is coded just after this event. For example, when the user is doing "classic thinking aloud", you code it at the same time when the first word is started, and end it when the evaluator asked a question, or when the user stopped doing the task and began to talk to the evaluator, or began to keep silence for a long time.

When doing the coding, since it is defined as "default state event" in Observer, which means when the next state event is chosen, the previous one will be stopped automatically. Hence, if the current state event is stopped, you just need to choose the next state event, and the previous one will be stopped. You do not need to choose "end" manually.

The state event means a period of time, not a time point, so it is ok with some other point events in the state event. As long as the state event lasts, then you do not need to code the state again. For example, for about 3 minutes, the user is talking his/her opinions of the images. It is state event "talk other than thinking aloud". Even though during this state event, there are many "affirmative response", "probing behaviour" by the evaluator or "negative comments" by the user, you do not need to choose the state event again. You just need to start the state event, and 3 minutes later end it and start another state event.

5. *Classic thinking aloud (Ericsson and Simon's level 1 and 2 data)*

Ericsson and Simon described three different levels of decreasingly reliable verbalization. Each level is characterized by the amount of interference caused by non-task related processing (Ericsson & Simon, 1993, p 79):

- The first level of verbalization is that which needs not to be transformed before being verbalized during task performance. When information is reproduced in the same form as it was attended, it is level 1 verbalization. At this level, there are no intermediate processes, and the participant does not need to expend special effort to communicate his/her thoughts, such as verbalizing numbers, since it is in the same form as they were originally encoded in STM.
- The second level of verbalization is that which needs to be recoded before being verbalized, such as verbalizing images or abstract concepts, but it does not bring new information into the focus of the participant's attention. "The recoding does not change the structure of the process for performing the main task." (Ericsson & Simon, 1993, p. 79)
- The third level of verbalization is that which requires additional cognitive processes to explain the thought process or thoughts, such as the explanation of thoughts, ideas, or hypotheses which require the participant to make intermediate inferences and change the normal flow of information in the STM. The information in level 3 is not attended by the participant at first, and does not belong to the normal course of task performance. Examples of additional cognitive processes include "making inferences about the subjects' own cognition, information retrieved from long-term memory at the researcher's request, and any outside influence, such as any comment or prompt from the researcher." (Boren & Ramey, 2000, p. 262).

6. *How to see whether it is classic thinking aloud (level 1 and 2) from the video?*
    - With the task performance: talking while doing the task. It is not explanation or talking feelings or ideas, but doing and talking at the same time.
    - Usually the speech by the user is "incontinuous"/ "discrete", such as "what does it say? En… I guess I have to copy…. And I have mine,… en… this one? No, let's see…."
    - Short words
    - Words repetition: some words may be talked again and again
    - Information in the short term memory, not from long term memory

    Usually there is no interrupts by the evaluator, except affirmative responses or "keep talking". If there are other responses from the evaluator, then the "thinking aloud" state event may be stopped.

7. *Talk other than classic thinking aloud data, such as comments, explanation, opinions by answering questions and so on*
    - Usually it is full sentence
    - Opinions, explanations, comments, answering questions, feelings
    - Information from long term memory

    There could be any responses from the evaluator. It is only stopped when they begin to be silence or begin to think aloud.

8. *Silence:* when nobody is talking for more than 10 seconds.

9. *How to code the point event?*

Code it when it happens. For point event, it is important to know how many times it happens, so if the point event is interrupted by the evaluator's response (such as affirmative responses, or probing behaviour) and then continue, still code the event once. In order to get a high reliability between the two coders, it is also important to know when to give the code. The codes are given when you just get the idea of what the point events are.

1) For the comments or suggestions from the user, or help behaviour from the evaluator, there is big chance to have communication in between. Just code the behaviour once if it has the same meaning and happens in one period of time.

    For example, when the user gave negative comments or suggestions, it lasted for a while, such as
    U: "if I knew which church the wedding gonna be",   (suggestion)
    E: yea   (affirmative)
    U: maybe it would be nice to have the picture of that church on the invitation (other)

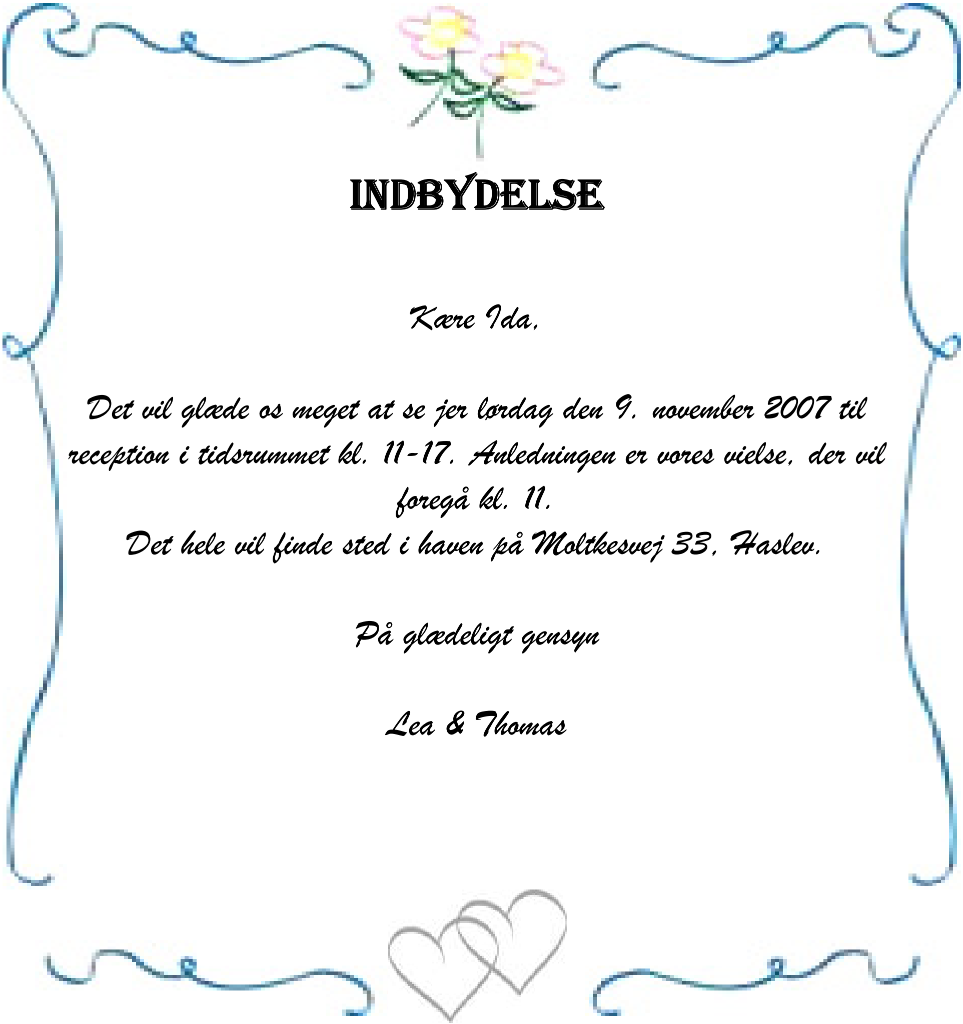
    Even though the comment is broken by the evaluator's affirmative response, it is still one comment, so just code it once, not necessary to code it twice.

2) For the verbal behaviour that could not last long, such as affirmative response or short reply, you code it every time when it happens.

    For example,
    U: I would not choose other people get married (negative comments)
    E: yea   (affirmative)
    U: and if I knew which church it is gonna be   (suggestion)
    E: yea   (affirmative)
    U: maybe it would be nice to have the picture of that church (other)

You code "yea" twice to show the evaluator is an actively listener. But of course, if the evaluator said "yea, yea" or "ok, ok" at the same time without any response by the user, then you just need to code it once. Such as the user said "I would not choose other people get married", the evaluator said "yes, sure", you just need to code it once.

Appendix 15: An example of a wedding invitation made by a Danish participant



INDBYDELSE

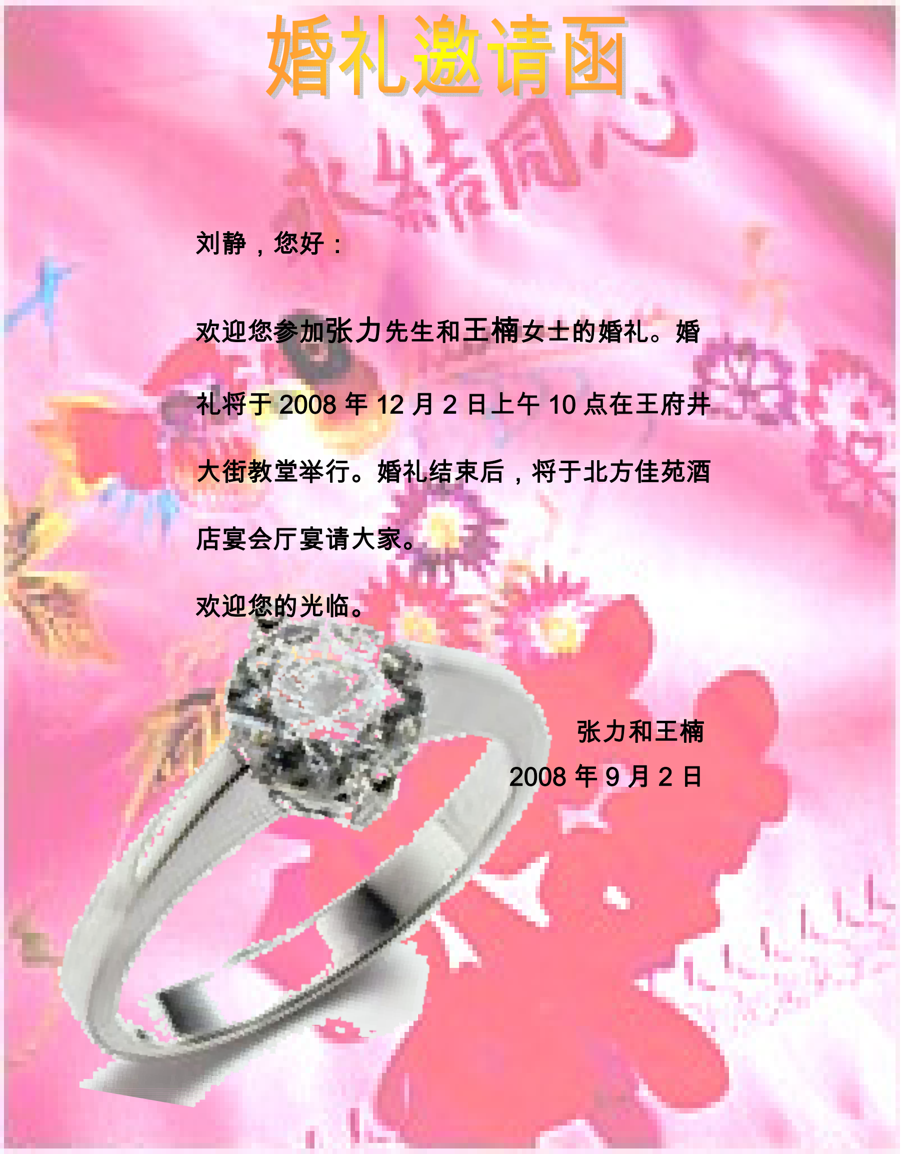
Kære Ida,

Det vil glæde os meget at se jer lørdag den 9. november 2007 til reception i tidsrummet kl. 11-17. Anledningen er vores vielse, der vil foregå kl. 11.
Det hele vil finde sted i haven på Moltkesvej 33, Haslev.

På glædeligt gensyn

Lea & Thomas

Appendix 16: An example of a wedding invitation made by a Chinese participant



婚礼邀请函

刘静，您好：

欢迎您参加张力先生和王楠女士的婚礼。婚礼将于2008年12月2日上午10点在王府井大街教堂举行。婚礼结束后，将于北方佳苑酒店宴会厅宴请大家。

欢迎您的光临。

张力和王楠
2008年9月2日

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or, How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning – A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus
– Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain
Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under 1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; ”Where process becomes the objective”

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk ’sted’ til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter.
En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος-tænkning hos Martin Heidegger

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetoder

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heteroge-

neous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the researcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance-effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt Studier i den biopolitiske ambivalens