A Field Study of the Relationship and Communication between Chinese Evaluators and Users in Thinking Aloud Usability Tests

Qingxin Shi
Copenhagen Business School
Department of Informatics
Howitzvej 60, DK-2000 Frederiksberg, Denmark
+45 72421536
Qs.inf@cbs.dk

ABSTRACT
Thinking aloud is the most widely applied usability evaluation method. In order to get reliable usability problems, it is necessary for the evaluators to establish a supportive relationship and communicate effectively with the users. This study investigated the relation and communication between the evaluators and test users in Chinese usability testing sessions. Field observations and interviews were conducted in five companies in Beijing. This research was based mainly on Nisbett's cultural theory and Boren and Ramey's thinking aloud model. The results of the study showed that Chinese users focused mainly on tasks, while evaluators focused on both users and tasks. Further, Chinese users did not think aloud actively; thus, in order to encourage users to speak out, effective communication skills were required for Chinese evaluators. Retrospective thinking aloud and explanation were also used in the tests. Finally, it discussed that communication was appropriate for the formative evaluations, but not for the summative evaluations.

Categories and Subject Descriptors
H5.2 [**Information interfaces and presentation (e.g., HCI)**]: User Interfaces--*Evaluation/methodology*.

General Terms
Human Factors

Keywords
Thinking Aloud Usability Testing, Culture, Formative Evaluation, Field Study

1. INTRODUCTION
Thinking aloud usability testing method has been extensively applied in industry to evaluate a system's prototypes of different levels of fidelity [1]. It requires representative users to talk aloud while performing a task or solving a problem. The primary goal of a usability test is finding a list of usability problems from evaluators' observations and analysis of users' verbal and non-verbal behavior. In order to make sure the users honestly disclose their thoughts and feelings, it is important for the evaluators to establish a trusting and supportive relationship with the users [2].

As Nielsen reports, the thinking aloud test is a typical method used for formative evaluation which is done in order to help improve the interface [3, 4]. In formative evaluation, the purpose is to find the good and bad aspects of the interface, thus communication with users tends to be necessary and important [5].

The most widely cited and comprehensive development of thinking aloud protocol has been issued by Ericsson and Simon [6]. This model emphasizes that when doing thinking aloud, there should be little interaction and communication between the user and evaluator. The only interaction may be asking the user to keep talking. Although Ericsson and Simon's thinking aloud model has been generally accepted in the field of cognitive psychology, like problem solving, in the usability research, usability professionals seldom conform to their model to do the test [7]. Considering the discrepancies between the observed thinking aloud testing and Ericsson and Simon's theoretical basis, Boren and Ramey [7] proposed a new approach based on speech communication theory. This study will be based on Boren and Ramey's thinking aloud model.

The requirement of establishing a supportive relationship and communicating effectively with the users may be one of the most important issues in formative thinking aloud usability tests in all the cultures, however, it is extraordinarily important for East Asian users. Because: 1) it is more difficult for East Asians to verbalize their thoughts because of the holistic thinking style [8]; 2) East Asians are socio-emotional relational orientation [9], which means people's effort and attention are focused on the interpersonal climate of the situation. It implies that relationship plays a more important role for East Asians than Westerners in the tests. This study investigates the relationship and communication in Chinese usability testing sessions based on field observation and interviews in Beijing. China is a big country with a variety of people. It is very hard to investigate usability tests in all the cities. In this study, we only did the field study in Beijing, because there are many kinds of IT companies and it is easier to get in contact. Besides, the users in the tests should be from the target user group, so no matter the users are from Shanghai or Beijing or some smaller cities, they will have similar characteristics for a specific test. Evaluators in different cities in China are also trained by the

"standard" way of doing the tests, which is from the Western countries. So the result of usability testing in Beijing, to an acceptable extent, could be generalized for China as a whole.

Recent work on the influence of culture on people's perception and thinking calls for a serious questioning of whether our knowledge and assumptions of the thinking aloud method are valid outside of Europe and the US [10]. Yeo's study [11] shows that culture impacts on usability testing. His research implies that in high power distance countries, in order to obtain an honest appraisal of a user interface, it is better to use experimenters who are in the same status or lower than the users. From his research, we get the inspirations that relations between the evaluators and users do play a role in usability testing, at least in some cultures. His research is based primarily on Hofstede's cultural theory, which assumes "the culture manifests at the behavioral level" [12, p 377-378]. Since thinking aloud usability testing is a cognitive activity [13], it may be more appropriate to use Nisbett's cultural theory, which discusses the cognition differences between East Asians and Westerners. This study will be mainly based on Nisbett's cultural theory.

Thinking aloud technique came from the Western countries. A study shows that the technique of speaking out thoughts while doing the task is more suitable for Westerners than East Asians. 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 [8]. East Asians believe that states of silence and introspection are considered beneficial for high levels of thinking [8]. Thus, asking East Asian users to speak out their thoughts while doing a task may influence their task performance more than it does so for the Westerners. Hence, it is worthy to investigate how to communicate in order to get all the users' feelings in the thinking aloud usability testing in East Asian countries.

In this study, first, the theories of culture and its influence on verbalization, communication and relation in thinking aloud usability testing are introduced. Then, communication and relation between the evaluators and users are discussed based on Boren and Ramey's thinking aloud model. Finally, the field study in China will be introduced and the data analysis will be discussed.

2. CULTURAL THEORY AND USABILITY TEST

2.1 Cognition and Thinking Difference

This paper is based on Nisbett's cultural theory [14, 15]. Nisbett's theory discusses the cognition and perception difference between Westerners and East Asians [15]. This theory is very relevant to usability tests since the thinking aloud usability evaluation method asks users to work on typical tasks and to verbalize their task performance and thought processes [16]. The whole process involves users' cognition and perception characteristics. The results of the usability test, i.e., usability problems which are found by the evaluators, are also involved in the evaluators' cognition and perception of the whole test process.

Based on Nisbett's cultural theory, in this study we define culture as an integrated pattern of local practice knowledge, cognition, and behaviour, which is both a result and process of interacting with the environment and society. A good usability evaluation method should fit the characteristics of the people in the target country.

Cultural analyses on cognition suggest that particular cultural practices tend to come up with particular modes of thinking in a given cultural context [8]. As mentioned, in Nisbett's theory, there are reliable differences in the modes of thinking between people from the East Asia and the West. In greater detail, Westerners' way of thinking is characterized by Nisbett as analytical, meaning that they tend to "think in a line." However, East Asians' way of thinking is seen to be more holistic in that they tend to "think in a circle." Kim [8] posits that thinking aloud is best suited to analytical 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 [8]. The reason may be that when East Asians, who tend to adopt holistic thinking, want to grasp the gestalt of the part, many elements will be held in thought at the same time. It will make the verbalization more difficult to do. In contrast, when Western people, who tend to adopt analytic thinking, break up the object into component elements it makes the verbalization easier to do [8].

Before doing thinking aloud usability testing in China, the Chinese usability practitioners need to consider whether the same thinking aloud method from the West is suitable, that is, without modification to 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. Considering the difficulty of verbalization for Chinese users, the evaluators need to resort to effective communication in order to get the users' real feelings and opinions about the application/ product.

2.2 Socio-Emotional Relational Orientation and Hong's Dynamic Constructivist Approach to Culture

In Nisbett's cultural theory, there are two types of relational orientation: task-focus orientation and socio-emotional orientation [9]. Task-focus orientation 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. Socio-emotional orientation means that people strive to maintain social harmony and their effort is focused on building a good relationship with others when they are doing the task.

As the study discussed above, Yeo [11] examined cultural factors that may affect results of usability evaluation techniques. Initial results showed that an important possible cultural factor was power distance: a test user who was of higher rank than the experimenter gave more negative comments about the product than a test user who was of lower rank than the experimenter. Based on Nisbett's cultural theory, people with a Malaysian cultural background tend to have a socio-emotional relational orientation. If users considered the evaluator to be of a higher rank, they would be more reluctant to provide negative comments [11] since they did not have a task-focus orientation and hoped to build a good relationship with the higher ranking evaluator. Thus, in Malaysian culture, in order to get honest results from usability testing, the experimenter needed to be of the same rank or lower than the test subjects. In Yeo's another research [17], he postulated that users needed to be familiar with evaluators before the users would talk honestly since they knew the evaluator's role

during the usability test, and knew that their negative comments would not destroy the good relationship with the evaluator.

Malaysia and China are both East Asian cultures. Chinese users may be similar to the Malaysian users, since China is also a typical socio-emotional relational orientation country. Chinese users may also be influenced by the evaluator's characteristics, status and behaviour, as the Malaysian users are. However, usability testing is a new technique in China and most tests are done in big cities where IT products or applications are investigated using educated people. From Hong's dynamic constructivist approach to culture [18], a culture's influence is not static, but related to the situation. The situation could also influence the effect of culture on cognition, affect and behaviour. In the situation of usability testing, the questions of whether Chinese users could be influenced by the perceptions of evaluators and how the Chinese users behave in the tests are unclear. In this study, we investigate how evaluators interact with users in order to have users honestly disclose their thoughts in the thinking aloud usability tests.

3. BOREN AND RAMEY'S THINKING ALOUD MODEL

Based on speech communication theory, Boren and Ramey [7] proposed a thinking aloud model for usability testing. This model focuses on the communication between the evaluator and test user. In the practice of usability testing, there are 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 [7, p.267]. The relationship and communication are much more important in this theory. The users do not ignore the evaluator. They expect a response, agreement, sympathy, etc., from the evaluator. In the speech communication model, the relation and communication between the evaluator and user are clarified as:

- 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 verbalizing thoughts fluently.
- When encountering contingencies during the usability test, interaction between evaluator and test user is required.
- In the practice area, it is permissible to probe with some questions to elicit more valuable information, which is not allowed in Ericsson and Simon's theory.

Ericsson and Simon's theory focuses primarily on cognitive processes. However, in the usability test, the main purpose is not only to get the user's cognition, but more importantly, to get the user's expectations, feelings, design ideas, etc., of the interface/software. So as long as the evaluator does not force his/her own opinion on the user, it is appropriate to communicate with the user to get more information about the interface. In order to achieve the good communication and suitable interaction, a warm, supportive and trusting relationship cannot be underestimated [19].

Boren and Ramey's work made a great contribution to usability research, but aimed mainly at proposing an alternative theoretical framework of thinking aloud usability testing and did not consider the tests in the East Asian culture. From the cultural theories discussed above, thinking aloud usability testing in East Asian cultures may need more interaction and communication because of the difficulty to verbalize the holistic way of thinking and the social-emotional relational orientation. Therefore, we carried out field studies in China to investigate in Chinese industrial area, how the evaluator builds the relationship and how he/she communicates with the test user in the usability tests.

4. METHOD

Field studies rely either on uncontrolled observation or on data collected for practical purposes [20, p56]. The data from a field study is "often more convincing than many experimental data for they describe real behaviour and its consequences rather than responses to hypothetical questions" [21, p85].

4.1 Companies Participating in the Study

In this field study, we observed six thinking aloud usability testing sessions in five companies in Beijing. Table 1 provides a summary of the companies. For the sake of anonymity, as we are not allowed to show the names of the companies, the names are replaced by letters.

Table 1. Information about the companies participating in the study

Company	Chinese or International Company	Clients of the company	Test sessions observed	The observed evaluators' experience (in years)
B	Chinese	Intern	1	1.5
M	International	Intern	1	1
S	International	Intern	1	2.5
U	International	Extern	1	2
X	International	Intern	2	6; 2

Four of the five companies are international companies and have been set up by Westerners. The fifth is a big Chinese IT company in China. Four companies in this sample perform mainly in-house usability testing since they are the usability departments or user-centered design departments in their companies. One company is a usability consulting company and

provides usability evaluations solely to clients outside the company. The evaluator's experience in the field study is from 1 to 6 years.

4.2 Data Collection and Analysis

In each company, we did field observation with video cameras of TA usability test sessions and afterwards interviewed the evaluators, test users and the usability department manager to get deeper information. The intention of the video recording was to produce a complete behavioural record of the evaluator and test users' relations and communications as they unfolded during the usability test session, including the time at which each instance of a behavioural relation occurred (events) or began and ended (states). The video observation and analysis were based on the seven foci in the "interaction analysis" framework [22]. The "interaction analysis" framework is one of the most popular frameworks for video observation, analysis and presentation, which was developed by Jordan and Henderson [22]. These foci are not the final analytical categories for understanding the interaction, but only a general framework for analytical categories which experience has shown to be valuable for video analysis [22]. The seven foci are: 1) structure of events, 2) temporal organization, 3) turn-taking, 4) participant structures, 5) problems and reparations, 6) organizing of the spatial activity, and 7) artefacts and documents. Table 2 gives an example of "turn-taking" to show how to use seven foci to record the observation.

Table 2. An example of "turn-taking" when recording the observation

3) Observation of "turn taking" (It encompasses the whole range of behaviours through which people can "take a turn", including turns at talk, turns with bodies and turns with artefacts)	**Explanations by the observer**
Turn taking was governed by the protocol. The evaluator gave instructions according to what was written on the protocol.	The evaluator seemed to be very fluent in using the protocol to ask the user questions.
E: What are you trying to look for? U: I am trying to find the main menu… E: how do you think the buttons here? U: Oh, I did not notice the middle button……	The user was not so talkative. The evaluator asked her questions to make her say something. The communication style was ask-answer.
E: what are you thinking? U: oh, I was trying to find the function of alarm clock….	The evaluator reminds the user to think aloud. The user preferred to give short retrospective reports.

There were three observers: two Chinese and a Dane. Since the Danish observer did not understand Chinese, he mainly recorded the behaviour and body language of the evaluator and user using the seven foci. Sometimes one of the Chinese observers translated some important speech for the Danish observer. We also briefly discussed the test immediately after each test. After all the observations were completed, the three observers discussed the notes together and elicited the main findings in a new form. In order to ensure the correct findings, the author analyzed the observations and interviews again through the grounded analysis approach [23] and focused on the relations and communications between evaluators and users in think aloud usability tests.

5. RESULTS AND DISCUSSION

The grounded theory model [23] for conducting a think aloud usability test was developed from the field studies presented in this paper, as shown in Figure 1. The goal of the analysis was not to have an accurate description of all data, but to form a quest for a conceptual theory abstract of the main elements in a thinking aloud usability testing.

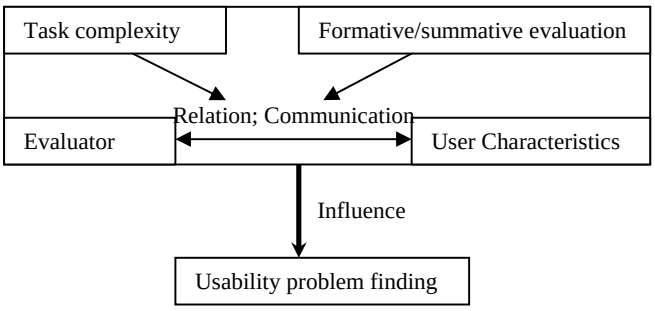

Figure 1. Theoretical model in a thinking aloud usability test.

Figure 1 shows the important elements in the thinking aloud usability test and their potential influence on usability problem finding. Previous studies have discussed the evaluator effect and user effect on usability problem detection [1, 3, 13], but there are few studies on the effect of interactions of the evaluator and user on usability problem findings.

As talked earlier, the primary goal of the usability testing is to find usability problems. In order to get the user's real feelings of the application/product, the evaluator needs to build a supportive relationship with the user to result in honest disclosure of thoughts.

Thinking aloud data which is described as reliable in Ericsson and Simon's model is not enough for a usability test [6]. Tamler [2] suggests that thinking aloud data which is generated by users themselves is often inadequate. The evaluator needs to use probe questions which are important for the interface but not noticed by the user[7]; beyond this, there is a need for an understanding of the user's speech and behaviour in order to get the user's real experience with the interface [2, 7]. Therefore, the communication and interaction is very important for the evaluators to find usability problems.

But, of course, how the evaluator communicates with the user is also related to the task complexity and evaluation types. In this field study, task complexity means a simple or complex task for the user, since from observation we could see that the communication between the evaluator and user shows different

Table 3. Overview of the results. N refers to the number of sessions in which a finding was made (6 sessions in total)

Main issues to be investigated	Main findings	N
Relations between evaluators and users	All users took the evaluator as an interviewer, facilitator, or guide. They seldom felt pressure or were uncomfortable with the evaluator.	6
	Evaluators always considered the user's feeling, e.g., "there is no right or wrong," "the test subject is the interface not you," "focus on both the task and the user during the test, and try to make the user relax."	6
	Users seldom considered the evaluator's feelings, and mainly focused on the tasks, and they also did not consider whether the evaluator understood them or not.	5
	The way the evaluators behaved in the usability testing was adjusted according to the users' personality, status and knowledge background.	4
	Evaluators thought the relations with the user should not be too close, which meant not using the very familiar users.	3
Communications	All evaluators used probing questions that they were interested in or wanted the user to explain.	6
	The users did not think aloud actively. Retrospective thinking aloud was very important for Chinese thinking aloud usability testing. For example, if the task was very simple, the user finished it quickly without saying anything, especially if the evaluator just watched his manipulation and did not ask him to think aloud. But after the task, the evaluator would ask the user how he/she had done that task.	6
	The users did not think aloud actively. Explanation was a big part in thinking aloud. If the task was very hard, most users also kept silent while thinking how to deal with it. At this time, the evaluators would ask questions or remind them to speak out. Then the users would explain to the evaluator what they wanted to do.	6
	Most evaluators in the TA usability testing did not really care about the users' thinking process, but cared more about their thoughts, feelings or opinions of the software/interface, e.g., the users seldom described how they did the tasks, but always gave their opinions of the interface.	6
	When the evaluators did not quite understand the users' meanings, usually they would repeat their understanding and ask the user for confirmation or correction.	6
	No training of thinking aloud. All evaluators just introduced thinking aloud to the users, and reminded them to speak out during the test, but none of them gave the user training.	6
	Evaluators seldom reminded the user to "keep thinking aloud," nor did they ask, "What are you thinking now?" Instead, they asked the users questions, such as "What are you looking for?" or "What are you trying to do?"	3

patterns with simple tasks and complex tasks. Further, whether the evaluator communicates with the user during the test also depends on the evaluation type, that is, whether it is a formative or summative evaluation. The two aspects will be discussed more in the following sections.

5.1 Relation and Communication

The data was analyzed using the seven foci. Then we organized the data from the observations and interviews into relations and communications between evaluators and users. Table 3 summarizes the two main issues and the main findings of each.

5.1.1 Relations between Evaluator and User

A trusting and supportive relationship is very necessary for a usability test [2]. Evaluators should clarify the purpose of the test, not be judgmental of the users' behaviour or speech, and encourage users to disclose their thoughts and feelings honestly without any worries. Most evaluators should be able to build this type of relation with the user in the practice area.

In the field study, both evaluators and users knew that the subject of the usability test was the interfaces, not the users. The users just took the evaluator as a facilitator, which helped them realize that there was no pressure placed on them. The results indicated that during the usability testing, most users just focused on the tasks, and seldom considered the evaluator's feelings, since they knew that their task was to help the evaluator find more usability problems, and their negative comments would not make the evaluator feel uncomfortable. This result is different from Yeo's research [11, 17] and also does not conform to the socio-emotional relational orientation theory. From the cultural theory, Chinese users were expected to consider the evaluator during the test. But it was not the case in this field study. In the field observation, we found that in two companies, the user almost never looked at the evaluator, not even when the user said something, e.g., when answering the questions from the evaluator, the user would still look at the computer screen. Some users did not care whether the evaluator understood their thinking aloud content. For example, in the interview after the test, a user said, "If she did not understand,

she would ask, so I did not think whether she understood or not when I was doing the task."

Reasons for this which were given during interviews with evaluators and users were [24, 25]: a) most users in the usability tests were well educated people, since the applications in the tests were IT products which were often used by those people; b) many people in Beijing were very familiar with being interviewed in many situations; c) the users knew the purpose of the interview since they were explicitly told that the test subject was the product, not them. Using Hong's dynamic constructivist approach to culture [18] as explanation, in the usability test situation, highly educated Chinese users who are living in big cities do not behave as most culture theories describe, which means that they focus mainly on the task and are not influenced by the evaluator's status, characteristics, or feelings. As we have discussed before, since usability testing is a new technique in China and is often carried out in big cities with the required target users, it is safe to say that the findings in this study can be generalized to the usability testing in China. In the usability test situation, Chinese users seem to be task-focus orientation, not socio-emotional orientation.

From the observations and interviews we got that the relationship between the evaluators and the users was 'student-tutor', similar to the 'learner-expert' relationship described by Boren and Ramey [7]. The evaluators considered themselves as the students, and regarded the users as the teachers from whom they were leaning the problems of the application.

During the tests, the evaluators always considered the user's feelings. The evaluator would adjust their facilitating behaviour according to a user's personality and background. For example, the evaluator in company S said, "For people in different status and with different features, I will try to make them feel equal. For expert, I will try to make them answer questions with patience. And for the user who is very talkative, I will try to make him/her focus on the topic. For people who do not like to talk, I will facilitate them to talk, and catch the basic information. For different people and different reaction, I will propose questions in a little different way. " Some evaluators also said, "If the user was a very familiar person in my real life, I would not think I could be the evaluator, since it would influence my behaviour, but if the users are children, maybe it is better to have a close relationship with them."

The main reason why users do not consider the evaluators but the evaluators always consider the users may be: the roles of users and evaluators are different. The users clearly know that their task is to help the evaluator find the problems of the interface. In order to fulfil the evaluators' needs, they will focus on the tasks. The evaluators' main task is to facilitate the test and help the users articulate their thoughts, so more problems can be found. To achieve this purpose, the evaluators have to consider the users' feelings and characteristics so that they are willing to verbalize their thoughts honestly and clearly.

5.1.2 Communication between Evaluator and User

In the usability testing, no evaluators passively listened to the user's "thinking aloud". They not only reminded users to keep talking, but more often used probe questions. This finding is the same as Boren and Ramey's work [7]. Considering the thinking aloud model proposed by Boren and Ramey based on speech communication theory [7], probing questions are necessary in a usability test, since users do not spontaneously address every important thing [2]. Evaluators need to use questions to call the user's attention to the important issues that they are not noticing. The evaluator in company X said, "We designed the protocol by ourselves, so we know what the important issues of this software are. I will probe questions that we are interested in if the user did not mention it." Nørgaard's research also found that evaluators asked questions about nonexistent parts of the system, speculative or hypothetical questions [26]. Questions that the evaluators asked during the tests "did not aim at understanding problems experienced by the users, but rather at encouraging users to predict possible problems" [26, p216].

From the observation, we found that the users often kept silent and did not speak out actively, so probing questions and reminding users to speak out tended to be very necessary in Chinese thinking aloud usability tests. This is confirmed in the cultural theory discussed above, indicating the cognition and thinking difference between East Asian and Western people, that is, Chinese people believe that states of silence and introspection are beneficial for high levels of thinking [8]. Verbalization of the thinking process will influence Chinese people's performance because of their holistic thinking style which is not easy to verbalize. In this study, Chinese users often forgot to speak out when they were doing the task. In order to benefit from their thoughts, the evaluator should interact with the user to get them to talk and complement the thinking aloud data.

Evaluators should use reminders and questions to make the users talk. From the field study, we found that the evaluators did not use reminders often; instead, they used questions. In the interview, the evaluators told us that when they could see what users were doing, they asked related questions, which would encourage the user to talk immediately. For example, if they knew that users were looking for something, they would ask, "what are you looking for?" and then they would tell them immediately about what they were looking for. The way of asking related questions to make people talk is more natural than asking people to "keep talking." But if the evaluator was uncertain about the users' behaviour, reminders of asking them to talk would be used in order to avoid misleading.

Communicating with users is very necessary in the usability test. Evaluators need to call the users' attention, clarify their real meaning and remind them to talk aloud in order to indicate how information is processed and to clarify the specific strategies that are used by users to complete tasks in a usability study [27]. Since Chinese users do not actively speak aloud, retrospective thinking aloud (RTA) and explanation tend to be major parts in the usability tests.

5.1.3 Task Complexity

In this field study, the tasks were divided into two categories: simple tasks and complex tasks. RTA was used more often in simple tasks and explanation was used more often in complex tasks. Generally, explanation is considered as a part of retrospective thinking aloud by many researchers [6, 27]. The reason we separate the explanation from retrospective thinking aloud is that we want to emphasize the breaks in the communication. RTA is usually used to collect the verbalization of a user's performance after the performance is over [27], which will not influence the user's task performance during the test; in other words, users will execute the task in a way that they usually do [27-29]. Explanation means that when users are

doing a complex task, they tend to keep silence and ponder it, and then explain what they are thinking or want to do under the requirement of the evaluator. The explanation in the current field study includes logic inference, perception explanation and strategy explanation [27].

From the observation, we got that although the tests were thinking aloud usability tests, the users often forgot to speak out. So the evaluators needed to ask questions or remind them to talk, but the timing of asking questions depended on the task complexity. If the task was very simple and users finished it quickly without saying anything, the evaluator would ask them how they had done that task after completing it. Then the users would give an RTA which described their behavior, comments and explanation. Since the task was simple, usually users did not give too much logic inference and high level explanation about their behavior. Describing the behavior and giving comments such as 'the task is easy' are more common for simple tasks. For example, in company U the task was a very simple search task and the user was asked to find articles in the website. It was very easy for her, so she completed the task very quickly (between 2 and 5 seconds per task) and she did not say anything during task performance. When she finished the task, she told the evaluator what she was doing and described her thought process during the task.

As discussed above, if the task was very difficult, most Chinese users kept silent while thinking how to deal with the task, which is representative of the holistic thinking style which is not easy to verbalize. In this situation, the evaluators would ask questions or remind them to speak out. Then the users would explain to the evaluator what they were thinking and wanted to do. The reasons we call it explanation not concurrent thinking aloud are: 1) the way the users spoke to the evaluator was not like thinking aloud. In the field study, the users would usually stop their executing behavior and explain to the evaluator, which was not like concurrent thinking aloud. Concurrent Think aloud describes users working on typical tasks while at the same time verbalizing what they are thinking and doing [27], which is not just speaking under the requirement of the evaluator without doing anything. 2) the content was not like the thinking aloud data described by Ericsson and Simon [6]. The users not only spoke the information in their short term memory, but also verbalized feelings, perceptions and strategy explanations, etc.

Here we discussed the relation between task complexity and retrospective thinking aloud and explanation. However, it does not mean the RTA always follows simple tasks and that explanation always follows complex tasks. It implicates then that in the practice of Chinese thinking aloud usability tests, in order to get the users' full view of the interface/product, the usability practitioners should sometimes combine RTA data and explanation with concurrent thinking aloud data.

Moreover, the think-aloud method in the usability practice is flexible. The evaluator does not actually require users to speak out all the time. For example, in our study if the task was very simple, sometimes the evaluators just watched users with no need of a reminder to speak out, because they did not need the description of the users' behavior. But if the task was difficult and the users kept silent because of thinking, the evaluators would remind them to speak out their thoughts. Guan, Lee et al. [27] found an interesting trend that subjects tended to produce more valid accounts and commit fewer fabrications in the complex tasks than in the simple tasks, suggesting that subjects put more thought into complex problem-solving and therefore could verbalize in greater detail.

5.2 Formative/ Summative Evaluation

Building a supportive relationship and communicating effectively with users is necessary in Chinese thinking aloud usability tests. But in some situations, evaluators may choose not to probe with questions so as not to disturb users at all. Whether evaluators interact with users depends on the evaluation types: formative evaluation or summative evaluation.

Formative evaluation is done in order to help improve the interface, and its goal is to identify a set of usability problems [3, 4]. 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 relies on quantitative metrics of effectiveness, efficiency, and satisfaction [3, 30, 31]. The thinking aloud method can be used for summative evaluation to get the user's cognition and thoughts of doing the task. But in the summative evaluation, the evaluator seldom communicates with the user during the test.

In the interviews after the tests, the evaluators told us that if using thinking aloud to do summative evaluation, such as recording the steps of completing a task in the mobile phone test, they seldom communicated with users when they were doing the task. But in formative evaluation, the purpose of which is to find the usability problems and positive aspects of the interface, communication with users tended to be a necessary and important part.

The above discussion implies that Ericsson and Simon's thinking aloud model [6] may be fit for the summative evaluation, whereas Boren and Ramey's thinking aloud model [7] may be suitable for the formative evaluation. Since thinking aloud is often used as a formative evaluation method in usability tests [4], Boren and Ramey's thinking aloud model which is based on communication theory seems to be more useful for usability testing.

The usability tests that we observed in this field study used formative evaluation. Besides cultural issues, the reason of why there were many communications and interactions during the tests could also be that the main purpose of the test was 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 needed to make sure that they got the users' real feelings and thoughts, which needed an effective communication and a good relation with the users [2, 7].

6. CONCLUSION

This paper has investigated how Chinese usability practitioners conduct thinking aloud usability testing in the industrial area in China through field observations and interviews, focusing on the discussion of communication and relation between the evaluators and users in the tests. From this field study, we learned that users mainly focus on the tasks and evaluators focus on both the task (facilitating the test and finding as many usability problems as possible) and the users. It is hard to ascertain whether this is only a phenomenon in Chinese usability testing situation or whether it is a worldwide phenomenon. What we can say is that in this field study of Chinese thinking aloud usability testing, the users did not show the social emotional

relational orientation as Nisbett's cultural theory discussed, but the evaluators did have this orientation. Because of the Chinese holistic thinking style which is not easy to verbalize in the higher levels of thinking, communication skills are required for Chinese evaluators, such as when to use reminders or when to use questions. Furthermore, the task complexity also influences the communication, such as when to use retrospective thinking aloud, when use explanation, or when there is no need to request speaking out, etc.

This study also discussed evaluators' interaction with users, depending on the evaluation style. The evaluators could only communicate with the users in a formative test, not in a summative test. Since thinking aloud is often used for formative evaluation [4, P170], communication becomes an important part in thinking aloud usability testing to find usability problems.

In this paper, we discussed only verbal communication. As Kim purports, "Indirect and nonverbal communication of meanings in conversations are more strongly assumed in East Asian cultural contexts than in European American cultural contexts" [8, P829]. We believe that in the next phase, nonverbal communication, such as body language or facial expressions will be analyzed [32]. In a forthcoming study, we will investigate and experiment with the types of relations and communications between evaluators and test users that are most effective for finding real usability problems.

7. ACKNOWLEDGMENTS

This study was co-funded by the Danish Council for Independent Research (DCIR) through its support of the Cultural Usability project. I thank my supervisor Torkil Clemmensen (Department of Informatics, Copenhagen Business School) who made a big contribution on the data collection. I also thank Huiyang Li and Xianghong Sun (Institute of psychology, Chinese Academy of Sciences) for the help of field observation and many other people who gave us access to the companies.

8. REFERENCES

[1] Law, E.L.-C. and E.T. Hvanneberg. Analysis of Combinatorial User Effects in International Usability Tests. in CHI 2004. Vienna, Austria

[2] Tamler, H., High-tech versus high-touch: The limits of automation in diagnostic usability testing. http://www.htamler.com/papers/techtouch/, 2001.

[3] Capra, M.G., Usability Problem Description and the Evaluator Effect in Usability Testing. 2006, Virginia Polytechnic Institute and State University: Blacksburg, Virginia. p. 292.

[4] Nielsen, J., Usability Engineering[M]. 1993: New Jersey: AP Professional.

[5] Shi, Q. and T. Clemmensen. Communication Patterns and Usability Problem Finding in Cross-Cultural Thinking Aloud Usability Testing. in CHI 2008. Florence, P: 2811-2816.

[6] Ericsson, K.A. and H.A. Simon, Protocol Analysis. Verbal reports as data. A Bradford Book. 1993: Cambridge Massachusetts.

[7] Boren, M.T. and J. Ramey, Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication, 2000. 43(3): p. 261-278.

[8] Kim, H.S., We talk, therefore we think? A cultural analysis of the effect of talking on thinking. journal of Personality and Social Psychology, 2002. 83(4): p. 828-842.

[9] Sanchez-Burks, J., R.E. Nisbett, and O. Ybarra, Cultural Styles, Relational Schemas and Prejudice Against Outgroups. 2000, University of Michigan.

[10] Clemmensen, T. and T. Plocher, The Cultural Usability (CULTUSAB) Project: Studies of Cultural Models in Psychological Usability Evaluation Methods, in Usability and Internationalization - Second International Conference on Usability and Internationalization, HCI International 2007, Beijing, China, Proceedings, Part I, N. Aykin, Editor. Springer: Heidelberg. p. 274-280.

[11] Yeo, A.W. Cultural Effects in Usability Assessment. in CHI 98, Doctoral Consortium. 1998.

[12] Faiola, A. and S.A. Matei, Cultural cognitive style and web design: Beyond a behavioral inquiry into computer-mediated communication. Journal of Computer-Mediated Communication, 2005. 11(1).

[13] Hertzum, M. and N.E. Jacobsen, The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction, 2001. 13(4): p. 421-443.

[14] Nisbett, R.E., The Geography of Thought. 2003, London: Nicholas Brealey Publishing.

[15] Nisbett, R.E. and T. Masuda, Cultural and point of view. PNAS 2003. 100(19): p. 11163-11170.

[16] Ramey, J., et al. Does Think Aloud Work? How Do We Know? in CHI 2006, April 22-27.

[17] Yeo, A.W. Global-software Development Lifecycle: An Exploratory Study. in CHI 2001.

[18] Hong, Y.-y. and L.M. Mallorie, A dynamic constructivist approach to culture: Lessons learned from personality psychology. Journal of Research in Personality, 2004. 38: p. 59–67.

[19] Shi, Q. and T. Clemmensen. Relationship Model in Cultural Usability Testing. in HCI International, 2007, Beijing.

[20] Festinger, L., & Katz, D., Research methods in the behavioral sciences. 1966, New York: Holt, Rinehart and Winston.

[21] Ross, N., Culture and cognition: implication for theory and method. 2004, Vanderbilt University: International Educational and Professional Publisher.

[22] Jordan, B. and A. Henderson, Interaction Analysis: Foundations and Practice. The Journal of the Learning Sciences, 1995. 4(1): p. 39-103.

[23] Punch, K.F., Introduction to social research-quantitative and qualitative approaches. 2005, London: SAGE Publications.

[24] Clemmensen, T. and Q. Shi. What is Part of a Usability Test? in CHI 2008. Florence, P: 3603-3608.

[25] Clemmensen, T., et al. Cultural Usability Tests - How Usability Tests Are Not the Same All over the World. in HCI International 2007. Beijing, China, P: 281-290 (Lecture Notes on Computer Science; 4559).

[26] Nørgaard, M. and K. Hornbæk. What do usability evaluators do in practice?: an explorative study of think-aloud testing in Proceedings of the 6th ACM conference on Designing Interactive systems, 209 - 218, 2006. USA.

[27] Guan, Z., et al. The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking. in CHI. 2006.

[28] Haak, M.J.v.d., M.D.T.d. Jong, and P.J. Schellens, Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. Interacting with Computers 2004. 16: p. 1153–1170.

[29] Haak, M.J.v.d., M.D.T.d. Jong, and P.J. Schellens, Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. BEHAVIOUR & INFORMATION TECHNOLOGY, 2003. 22(5): p. 339–351.

[30] Barnum, C.M., Usability testing and research. 2002: Longman.

[31] Leventhal, L.M. and J.A. Barnes, Usability Engineering: process, products, and examples. 2007, Upper Saddle River, New Jersey: Pearson Education, Inc.

[32] Yammiyavar, P., T. Clemmensen, and J. Kumar. Analyzing non-verbal cues in Usabiliy Evaluation Tests. in HCI International, 2007, Beijing.