Cultural Usability Tests – How Usability Tests Are Not the Same All over the World

Torkil Clemmensen[1], Qingxin Shi[1], Jyoti Kumar[2], Huiyang Li[3], Xianghong Sun[3], and Pradeep Yammiyavar[2]

[1] Department of Informatics, Copenhagen Business School, Denmark
{tc.inf,qs.inf}@cbs.dk
[2] Indian Institute of Technology Guwahati, Assam, India
{jyoti.k,pradeep}@iitg.ernet.in
[3] Inst. of psychology, Chinese Academy of Science, Beijing, China
{lihy,sunxh}@psych.ac.cn

Abstract. The cultural diversity of users of technology challenges our methods for usability evaluation. In this paper we report on a multi-site, cross-cultural grounded theory field study of think aloud testing in seven companies in three countries (Denmark, China and India). The theoretical model that emerges from the data suggests that the production of a usability problem list is multi-causal and subject to cultural variations. Even the way usability problems are experienced by test participants may be different. In the discussion we outline practical guidelines for a test that is more sensitive towards cultural usability.

Keywords: Usability test, think aloud, cultural usability, field study.

1 Introduction

Culture plays an increasing role in discussions of information and communication technology. As of today, we do not have any formal methods to guide us in evaluating a product to a certain standard while being sensitive to cultural issues. Cultural usability tests are not yet established methods. Thus, in this paper our point of departure is a look at the methods that we have and to consider the following research questions. Which current practices of think aloud (TA) usability testing address cultural diversity for both the evaluator and the user? How do test users respond to the test instructions and test methodology? Which interfaces are more influenced by cultural diversity and current usability test practices? We try to understand the most effective way to get usability feedback from test users' without actually disguising the usability problems.

The rest of this paper is organized as follows. The following section deals with the multi-site cross-cultural method, section three presents the theoretical model that emerged from the observations, and section four discusses our findings.

2 Method

We approached the research questions by studying current TA usability testing practice in companies that test software and interactive products for the local market

N. Aykin (Ed.): Usability and Internationalization, Part I, HCII 2007, LNCS 4559, pp. 281–290, 2007.

in three countries (China, Denmark, and India). The selection of participants was done on the basis of global ethnography [3] which suggests that we need multiple studies in different settings to shed light on a phenomena. We used a site-based procedure [1] for locating and recruiting qualitative study participants to learn about each cultural setting and to gain entry into the setting. Our procedure had five steps:

1. Specification of the characteristics relevant to the sample: geographic (Denmark, India, China), socio-cultural (experienced moderators/evaluators and local test users taking part in TA tests), company characteristics (we only used companies that did professional user testing services of software products to the local market)
2. Generation of a list of sites - the places where TA tests were done
3. Estimate of the composition of clientele at each site by contacting a 'gatekeeper' for each site and asking for appropriate statistics for the site and helping to gain entrée; the 'gatekeeper(s)' in our case were the managers in each company who had the daily responsibility for running TA tests'
4. Recruitment of participants and 'gatekeeper' (in reality, the manager) and an agreement on when and how the TA tests at the site could be observed and the test users and moderators could be interviewed
5. Recruitment of individuals from sites and maintenance of a table indicating the characteristics of the participants in the sample [13] to help the researcher to assess the quality of recruitment. In our case it revealed that we had to live with demographic between-site differences between test users and moderators

We achieved a sampling diversity that was saturated in the sense of: a) we were not able to get into contact with more willing TA test vendors; b) we did achieve a reasonable amount of variation in our sample; and 3) it was quite clear that the three geographical categories of TA test vendors were clearly independent (none of the companies cooperated). In each company, we did field observation with video cameras of TA usability test sessions and afterwards interviewed the evaluators and the test users. We were three observers/interviewers: an Indian, a Danish and a Chinese. We made sure that two or more were present at all observations to increase cultural validity. We made it explicit that the TA test was to be run appropriately; i.e., the usability test manager should make the decisions about how to run the usability test to allow us to observe the current practice of TA in that company. The test application should be aimed at the local market. If it was not possible to observe a customer-paid TA test, we asked the company to redo a recently run customer-paid test. This procedure gave us a total of 52 hours of observation in three languages across seven companies. The analysis of the interviews was done through a grounded analysis approach [7] in which we focused on the production of usability problems in the conduct of think aloud usability tests.

3 Results

The grounded theory model for conducting a think aloud usability test was based on previous conceptual and empirical work within the cultural usability project [4], and then developed further 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 a quest for a conceptual theory abstract of time, place and people [7].

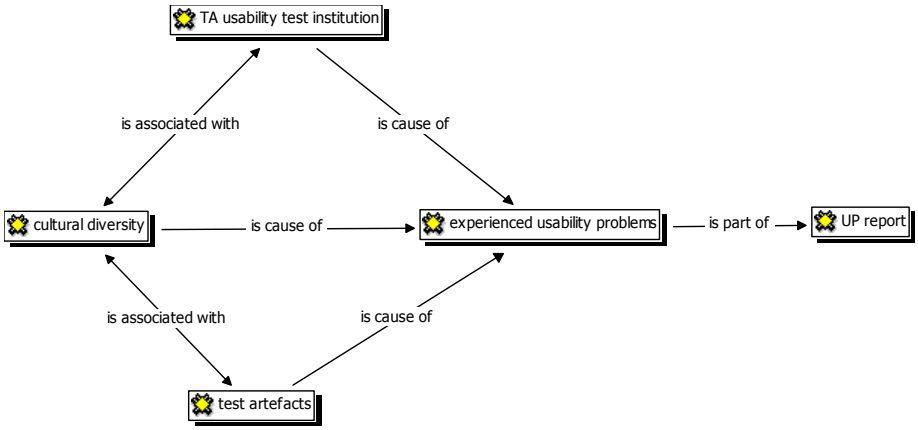

Fig. 1. Theoretical model for conducting think aloud usability tests

3.1 Cultural Diversity in Moderators and Test Users

The two traditional independent variables of: 1) age and 2) gender emerged from the field data as important parts of the construct of cultural diversity among moderators and test users. The Indian moderators saw age and gender differences between moderator and test users as no problem with urban users, albeit as a potential problem with rural users of a traditional cultural background where it was important to "speak their language", "be polite and respectful towards the elderly", and "be keen to listen". Specifically, a senior moderator could frighten a young test user with a traditional cultural background and require the moderator to "go an extra mile to communicate". Even in a usability test conducted remotely in which the moderator would not see the test user, the moderators would want to know the age of their test users. Additionally, if the user were a female from a traditional family, a male member of her family had to be present during the pre-interviews and the test, or the moderator needed to be a woman, (moderator): "It takes a woman to interview female users with traditional background (female members of a traditional family)". In the Chinese tests, age and gender differences were emphasized, (moderator): "….the way they do the test will be different…". "…if the user is male it is better if the moderator is a female…". In the Danish tests, the participants were positive towards the test users who were most similar to them in terms of age, gender and job experience (audience): "He was a very good test user", and hostile to the extent of being offensive towards a test user who was different in age and gender (audience): "She is an old …..."; at one point the moderator could hear the audience laughing and he was afraid that the user should (moderator): '…play this…'.

3.2 The Usability Test as an Institution

The cultural diversity of test participants explains partly the usability test practice observed in the three countries in terms of variations and additional properties of established usability tests [2]: 1) specific test goals and concerns, 2) real work tasks to be done, 3) think aloud procedures, 4) test users, evaluators, observers, clients/managers,

designers, in-house trainees, 5) recovery procedures, 6) evaluator room and observer room, 7) test applications and usability problem lists/reports. All incidents of usability tests that we observed did share this general approach to usability testing, and most exhibited interesting variations of the properties of the test.

Specific test goals and concerns. In the Danish tests, one goal was to get the test users to think aloud to allow insights into the test user's cognitive work with the test application. However, from the beginning it was unclear exactly when the user was required to think aloud. The evaluator asked about the user's opinion and preferences most of the first part of the two hours that each test took, and then gave the user a more specific task to do towards the end that was more like think aloud. In the Chinese tests, it was also a goal to do a think aloud test. All the moderators said it was a think aloud test, although some (moderator B + S) pointed out that in their case the test was actually a combined think aloud and interview. Moderators asked questions and the users answered throughout the whole session in all the studied companies. Furthermore, upon occasion the users did the tasks silently, and then after the task completion the moderator probed or asked for explanation, i.e., no think aloud occurred. The reason for this non-interference from moderators was given as (moderator M): "… there is no need to disturb the user when she takes the correct steps….". The user did not seem to think aloud actively in the Chinese test. In the Indian tests, it was an explicit goal with the scenario that the test user should be open and tell about his or her mind. However, all moderators were obviously concerned that usability testing should be seen as part of a user centered design process, (Moderator 3): "Basically the final outcome for us has to be a tangible physical design. So for that we get more detailed information…".

Real work tasks to be done. In the Indian tests, all moderators used a detailed protocol to get the users started on different tasks, (Moderator): "Through scenarios we start the task, when the user has stopped talking and he has answered most to the thing and if I have no questions to ask, the task is stopped", (Moderator): "I just tell them this is the whole scenario, this is what we have to do…to stop: The person has finished all tasks…", (Moderator): "I say: this is what to do, how will you do it…To end: it is a natural end, when the final step has been reached…". It was a concern that the users understood their tasks, which they did (user 1): "I think it is a good task", (user 3): "It was a good task, it made me realize, I found it quite interesting, because I go on sites all the time- I was able to compare with other sites….makes for a healthy ..comparison". In the Chinese tests, task management was informal, both regarding the start and end of the user's task, (moderator): "…lets see the next task…", or simply "After I introduce a task, then it will begin". The concern of having a detailed protocol to guide the test was not always strong, (moderator B) "…if you have many protocols it will scare the user, make the user very nervous, uncomfortable", so that the argument would be that if a moderator needs a protocol, one should ask her or him (moderator B) "Did she do the real usability test or not? In the real usability test, they should not have….it would scare the user...". Another moderator was quite concerned with having and following a protocol, (moderator S): "…sometimes I read out the task in order to make all users get the same instructions…". In the Danish tests, there was a countless number of steps in the protocol, which was partly due to the client wanting a large part of the test application (a website) tested in one long test.

Think aloud procedures. In the India tests, users were actively thinking aloud and speaking out, (User 1 repeatedly) "Now I will do...", and also sometimes thinking in phrases: "I think in today's age…". When necessary, the moderator used reminders to help the user to think aloud. Reminders could have many forms, see Table 1.

Table 1. Reminders to think aloud, from usability tests in Indian company

– please think aloud	– can you say what you are finding
– what are you looking for in this page	– talk to me what are you looking at
– what is happening now	– keep talking
– whatever you like or dislike or you think you can say	– could you tell me more information about what are you doing, you are expecting…

All moderators used hands and arms to make lively gestures to support their speech. The moderators were trained in body language, (manager): "… moderators are trained about the body language…it is a matter of practice that some follow it and some are in the learning phase. The idea is to keep the user comfortable at all times – by communication, body language, other settings etc. …". From the India tests emerged several ways of probing for more information: 1) when the moderator reacts to the users' initiative, e.g., (moderator) "You just said right", (user) " Yes yes I mean this thing…"; 2) when the moderator gives direction to the user as when the moderator asks, "Did you notice that?" and shows something on the screen, and the user says: "No that was my mistake I didn't", 3) when the moderator actively wants to help the users: "Usually in the first task I always ask such questions in order to let the user know what is his job, later I don't do that because then the user would know what he should do", and 4) when the moderator actively wants to dig deeper into the users think aloud, e.g., when a user says "I am not happy with the description…", and the moderator asks: "Why is that?". The difference between asking questions, reminders to think aloud, and probing is not necessarily clear in usability test practice.

In the Chinese tests, reminders to think aloud were used, (moderator UC): "You have finished the task, what did you think just now?" This kind of retrospective reminders appeared to be necessary. In company (UC) 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. In company B, when the user tried to edit (input some words) and was silent for maybe more than one minute, the moderator simply wrote something down on the paper and never reminded the user to think aloud. After a long time, the user said, "I have never done it before….". In company (M), the user did not think aloud at the beginning of the first task (see who called you), and then the moderator reminded the user two times: "…you can talk when you are doing…Which input method are you using now?", and only then did the user begin to think aloud: "…oh I find it, I should fix it, and click the button….". Paradoxically, while almost all users did not consider if the moderator understood them or not, many times after having been silent for some period, they explained to the moderator. The think aloud was an explanation, i.e., the think aloud was retrospective think aloud [5].

Participants. In the Danish tests, the audience consisted of two designers, two marketing people and two managers from the client company. They were obviously important: the senior usability specialist in the company (the head of the usability company) was observing the whole test and developing the usability report in the form of one powerpoint show for each user concurrently with the TA test in front of the audience. Part of the reason for this was to maintain client relations. In the Indian tests, client relations were also important, but the focus was on the relation between the moderator and user, (manager): "The relationship between the moderator and test user should be that of a teacher and a student, the moderator should be the student and the user the teacher (master and apprentice, actually)". The user was supposed to take the role of a design critique who, under guidance of the moderator, evaluates the website. In the Chinese tests, the focus was on the test user-computer dialogue: In two companies, the user did not at all seem to consider the moderator, did almost never look at the moderator, not even when the user said something, e.g., when answering questions from the moderator, the user would still look at the computer screen. Almost all users focused on the task, thinking that the moderator was just a person who gave them instruction and facilitated the test, and they considered him "… just a little more than if I was answering a questionnaire from anonymous sender…". Reasons for this were given during interviews with users to be a) many people in Beijing are very familiar with being interviewed in many situations, b) all users are well educated, know the purpose of the interview, know they are not the subject, c) the users are explicitly told that the test is about the product, not themselves.

Recovery Procedures. A concern in all companies was that the user should not be stuck unnecessarily in a task. For example in the India tests, the moderator helped if necessary, (User 4): "The two times I got stuck he helped me…".

Evaluator Room and Observer Room. In the Danish tests, an observer room accommodated up to 10 clients, observers, trainees, researchers, who through a one-way screen could watch the evaluator/moderator and the test user in the evaluator room. A video of the test user, a PC screen capture and sound were played on monitors in the observer room. A note-taker in the observer room took notes, while the moderator in the evaluator room did not take notes during the session. In the Indian and Chinese tests, there were arrangements very similar to the Danish with a separate note taker in the observer room. In China, Company B was different from the other companies in the sense that there was no dedicated note-taker, the moderator was simultaneously note-taker and wrote the report afterwards. Uncharacteristically, in Company NN the moderator was in the observer room, looking through the one-way screen and interacting through a button-operated telecom with the test user who was alone in the evaluator room.

The Usability Problem Lists/Reports. In the Danish company, the note-taker had the main responsibility for the final usability report to the client. During the tests, he was writing the draft report directly into MS power point, compiling the different users' responses, and structuring the report according to the order in which the test application areas (web site page) were being tested. The moderator was consulted before the note-taker finalized the report to the external client. In the Chinese company B and NN the moderators used a template to write the test report and presented it to the in-house clients. In the case of external clients, a full report was

made. In the Indian company, usually the note-taker used a predefined MS excel template with the test application areas as stated in the test protocol listed in a row and the users in a column, ready to enter the users' reactions, performance scores and satisfaction measures directly into the spreadsheet and then later produce the report. The report usually consisted of 50-80 pages document and a power point slide show.

3.3 The Test Application

What emerged from the observations was that the test users' work with the test application was efficient when they had already been primed with the typical functions of the test application, but were less efficient when they had not been primed. This supports findings from basic research in psychology and anthropology that subjects become fixed on the design function of the object after being exposed to a demonstration of the object's function [6]. It supports the findings from [12] that novice computer users from different cultural groups are not necessarily comparable, but can be seen as relative novices compared to expert computer users with similar cultural backgrounds. Being a novice user in a culture which surrounds you with computers is not comparable to being a novice user in a culture with few computers.

We observed that test users in usability tests will often be urban, modern, young, with higher education, be fluent in English and with substantial computer experience. In the Indian tests, the artifact was a newspapers website and the test users were appropriate users of online news sites: urban, young (20-30 years) with higher education, fluent in English and with experience in using this type of website. In the Danish tests, the test application was a website for an internet and telecommunications provider. The users had higher education, were 25-55 years of age, and end-user of medium to professional expert users of the test application. They spoke their local language, but were also proficient in English. In the Chinese tests, the test application in company UC was an e-learning public school website and the test user a female user, young, not fluent in English, but had experience with similar applications. In company B the test application was a search engine website, and the user was a male user with higher education, not fluent in English, had considerable experience in the test application but not in the tested new functions. In company S the test application was Web-based chat software; in company M Mobile phone interfaces (pen vs key) and in both cases the users were young, female, highly educated, fluent in English and had extensive knowledge of similar applications as test application. In company NN the test applications were: a Web-based work flow tool, with a male user in his thirties, had higher education, was an in-house employee with experience in similar systems, was English speaking, and had a Mobile phone service provider's customer website with a test user who was young and had experience with similar applications. Obviously, in current practice of TA usability tests there is awareness of recruiting users that culturally meet the test applications affordances. One important observation, however, is that when new software is tested, users are not always sufficiently fixed on the design function of the software, i.e., even if the user knows the kind of application being tested, he or she may have little clue about the intended use of the specific functions being tested – while other users may have a clear idea about this. This variation may be one cause of the variations in experienced usability problems.

Finally, some test applications are made for other user groups and require users with other backgrounds like elderly/children and/or rural, traditional, strongly religious, low education, local language only and no computer experience. Furthermore, they may have no or very few computers in their daily environment and hence little general knowledge about what computers can do.

3.4 Experienced Usability Problems and Usability Report

The cultural diversity in moderators and test users, the usability test practice and the affordances of the test application together contribute to the experienced usability problems and to the final usability report to the client. This multi-causal model of experienced usability problems partly contradicts existing usability problem theory that says that the list of usability problems only reflects the properties of the product being tested [11]. However, our study supports the findings that cultural diversity in the test users and moderators [10] and variations in the test setup [9] influence the detection of usability problems. In the Indian and Chinese tests, usability problems were experienced as interactions between test users and moderators, i.e., as co-discovery episodes. In the Indian tests, users when asked, could come up with one or two major problems they believed to have found during the tests, for example (User 2): "…he [the moderator] did give me indications….like where would you go to search for flights….". In the Chinese tests, the user also suggested design changes. For example a moderator asked the user "…why did you not find it right now…", then the user said "…oh I didn't notice this part…", then the moderator said: "…that means they should not put this in this part?" then the user said "…yes they should put it in flash or something…". In Danish tests, the users were quite confident that they could identify usability problems by themselves; the users did not refer to interactions with the test moderator, but instead focused on their own needs as users. These differences could mean the moderator-test user relation extends beyond the session and into the post-session period of fixing the usability problem list.

4 Discussion

Compared to previous studies of cultural usability and usability testing in natural settings, this study is distinctive in its multi–site, cross-cultural approach. A theoretical model of the production of usability problems in seven usability test vendors across three countries was constructed through a combination of interaction analysis observation and grounded theory analysis, which included systematic use of observers with different cultural backgrounds, and checking concepts with the participants to increase validity. This model encompassed cultural diversity in test users and moderators, variations in the conduct of usability test, and assessment of user-technology-fit in a conceptual framework for appreciating nuances in the outcome of usability tests in different regions of the world.

The cultural diversity in the background of users (and moderators) suggests that all usability is culturally specific and concrete. As Honold [8] observed, cultural orientation manifests itself in artifacts (technical products) and institutions (organization). The female, elderly Indian user of German washing machines who meets a male, young usability professional from Germany may not reveal the

subtleties of her preferences for quick wash programs (she wants morning tasks to be finished by noon (app. 3 hours to do the work), top-served machines (she can control the water level in an environment with less water), and similar highly contextual issues. In relation to usability, age and gender issues have to be considered together with the objective of users, environment, infrastructure, division of labor, organization of work, mental models based on previous experience and tools to understand the nature of the usability problem list [8]. The users' cultural background may be even more complicated than age and gender issues suggest. One example is when the availability of Arabic language interface in Gmail ™ gives an bicultural (north African and Danish) middle-aged occasional user a good feeling, even if he does not at the moment apply the interface to write an Arabic mail to some of his Arabic speaking friends, and even if the possibility of writing in Arabic is not accessible – just because the availability of an Arabic interface to the email system indicates the possibility of accommodating Arabic email dialogs. His daughter, who is a power user of Gmail, may however, experience that when she changes from Danish left-to-right to Arabic right-to-left language interface the functions switch place and she is reduced to a novice user. Switching to a culturally different interface may ruin a user's memorized information structure.

We offer the following recommendations in order to avoid producing biased lists of usability problems when doing user tests, especially cross cultural tests:

1. Balance out potential "hidden user groups" within user segments, for example, users who adapt quickly to international test conditions (used to foreigners) vs users who do not (are not used to foreigners), and culturally sensitive (traditional, rural) vs not sensitive (modern, urban) users.
2. Calculate the detection rate for each "hidden user group" [10]. To avoid missing critical usability problems, pick evaluators from an evaluator group suitable to the "hidden user groups" (calculate the evaluator effect).
3. Have different versions of your test protocol ready that include different types of scenarios, such as Bollywood dramatic (India closed users) vs traditional (China, Denmark, India open users), and/or different probing questions such as direct and frank (Chinese) vs. indirect (Denmark, India).
4. When writing up the report, have different templates for different clients like foreign clients vs. home market clients.
5. Plan to repeat tests in China after 5 years, because target users change very quickly, although this may not be necessarily true in India or Denmark.

Acknowledgments. This study was co-funded by the Danish Council for Independent Research (DCIR) through its support of the Cultural Usability project. A big thank to Thomas Plocher, Honeywell, and Apala Chavan, Human Factors International India, and to the others who gave us access and helped us in the companies.

References

1. Arcury, T.A, Quandt, S.A.: Participant Recruitment For Qualitative Research: A Site-Based Approach To Community Research In Complex Societies. human Organization 58(2), 128–133 (1999)
2. Barnum, C.M.: Usability testing and research. Longman, New York (2002)

3. Burawoy, M.: Global Etnography. California Press (2000)
4. Clemmensen, T., Plocher, T.: The Cultural Usability Project (CULTUSAB): Studies of Cultural Models in Psychological Usability Evaluation Methods, Invited contribution to a parallel session. In: HCI International, (Beijing, July 25-27, 2007)
5. Ericsson, K.A., Simon, H.A.: Protocol Analysis. Verbal reports as data. Cambridge Massachusetts (1993)
6. German, T.P., Barrett, H.C.: Functional Fixedness in a Technologically Sparse Culture. Psychological Science 16(1), 1–5 (2006)
7. Glaser, B.G.w.t.a.o.J.H.: Remodeling Grounded theory [80 paragraphs] Forum Qualitative Sozialforschung / Forum: Qualitative Social reseach [Online Journal] (March 4, 2004)
8. Honold, P.: 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 (2000)
9. Kjeldskov, J., Skov, M.B., Als, B.S., Hoegh, R.T.: Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In: Brewster, S., Dunlop, M.D. (eds.) MobileHCI 2004. LNCS, vol. 3160, pp. 61–73. Springer, Heidelberg (2004)
10. Law, E.L.-C., Hvanneberg, E.T.: Analysis of Combinatorial User Effects in International Usability Tests. in CHI (Vienna, Austria 2004), pp. 9–16 (2004)
11. Preece, J., Rogers, Y., Sharp, H.: Interaction Design: Beyond Human-Computer Interaction. John Wiley & Sons, New York (2007)
12. Rau, P.-L.P., Choong, Y.-Y., Salvendy, G.: A cross cultural study on knowledge representation and structure in human computer interfaces. International Journal of Industrial Ergonomics 34(2), 117 (2004)
13. Trost, J.E.: Statistically nonrepresentative Stratified Sampling: A Sampling Technique for Qualitative Studies. Qualitative Sociology 9(1), 54–57 (1986)