International Journal of Human-Computer Interaction
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653655

Templates for Cross-Cultural and Culturally Specific Usability Testing: Results From Field Studies and Ethnographic Interviewing in Three Countries

Torkil Clemmensen[a]
[a] Copenhagen Business School, Denmark

Accepted uncorrected manuscript posted online: 04 March 2011

Online publication date: 17 May 2011

To cite this Article Clemmensen, Torkil(2011) 'Templates for Cross-Cultural and Culturally Specific Usability Testing: Results From Field Studies and Ethnographic Interviewing in Three Countries', International Journal of Human-Computer Interaction, 27: 7, 634 — 669, doi: 10.1080/10447318.2011.555303, First posted on: 04 March 2011 (iFirst)
To link to this Article: DOI: 10.1080/10447318.2011.555303
URL: http://dx.doi.org/10.1080/10447318.2011.555303

INTL. JOURNAL OF HUMAN–COMPUTER INTERACTION, *27*(7), 634–669, 2011

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2011.555303

Taylor & Francis
Taylor & Francis Group

Templates for Cross-Cultural and Culturally Specific Usability Testing: Results From Field Studies and Ethnographic Interviewing in Three Countries

Torkil Clemmensen
Copenhagen Business School, Denmark

The cultural diversity of users of technology challenges our methods for usability testing. This article suggests templates for cross-culturally and culturally specific usability testing, based on studies of usability testing in companies in Mumbai, Beijing, and Copenhagen. Study 1 was a cross-cultural field study of think-aloud testing done by usability vendor companies in the three countries. The result was a grounded theory of cultural variations in the production of a usability problem list. Study 2 was a follow-up, ethnographic interview study of how the companies typically perform usability tests. The result was the construction of templates for usability testing. The culturally specific templates were in Mumbai "user-centered evaluation," Copenhagen "client-centered evaluation," and Beijing "evaluator-centered evaluation." The findings are compared with related research, and the implications are pointed out. The templates can be seen as a simple and practical way to plan, compare, and improve the way usability testing is carried out in multiple, different cultures and countries.

1. INTRODUCTION

Culture plays an increasing role in the use of information and communication technology. One example is when the availability of an Arabic language interface in a webmail system gives a bicultural (North African and Danish) middle-aged occasional user a good feeling, even if he will not at the moment use the interface to write to some of his Arabic-speaking friends, and even if the possibility of writing in Arabic is not accessible, but just because the availability of an Arabic interface indicates the possibility for applicable and appropriate Arabic e-mail dialogs. His daughter, who is an expert user of the webmail system in Danish

This study was supported by a grant from the Danish Research Council for Independent Research to the Cultural Usability project. Thanks to the anonymous individuals and the managers from the companies who agreed to take a lot of time out of their busy schedules to participate in this study. Thanks to Qingxin Shi and Jyoti Kumar for collaboration on Study 1, and to Pradeep Yammiyavar and Xianghong Sun for helping with access to companies in India and China.

Correspondence should be addressed to Torkil Clemmensen, Department of Informatics, Copenhagen Business School, Howitzvej 60, 5.15, DK – 2000, Frederiksberg, Denmark. E-mail: tc.inf@cbs.dk

left-to-right writing mode, is, on the other hand, reduced to a novice user when she switches to Arabic right-to-left writing mode, because the webmail interface functions also switch from left to right. Her memorized information structure becomes an obstacle—useless when she interacts with a culturally different interface.

Despite the influence of cultural variations on human–computer interaction (HCI), we do not have sufficient methods to guide us in evaluating a product to a certain standard while being sensitive to cultural issues. In the past, researchers have suggested paradigms for culture-specific HCI such as "cultural computing" (Rauterberg, 2006), "culturally sensitive information technology (IT)" (Zakaria, Stanton, & Sarkar-Barney, 2003), "intercultural usability engineering" (Beu, Honold, & Yuan, 2000), "culturability" (Choi, Lee, & Kim, 2006), and "cultural usability" (Barber & Badre, 1998; Hertzum, 2010). Attempts have been made to include cultural knowledge such as cultural dimensions (Marcus & Gould, 2000), cultural factors (Smith, Dunckley, French, Minocha, & Chang, 2004), cultural constraints (Norman, 1988), and cultural models (Jagne & Smith-Atakan, 2006), in research into HCI in general, and into cultural usability specifically (Clemmensen & Plocher, 2007; H. Sun, 2004). A major finding from the existing literature on culture and HCI is that there are cultural differences in the models that different user groups have of what HCI is. For example, Chinese users adapt a more holistic model of what it is to use software, compared to European users (Smith et al., 2004). The cultural differences imply the need for localization of the software design (Marcus & Gould, 2000) and for localization and cultural adaptation of established user experience and usability evaluation methods (Clemmensen, Hertzum, Hornbaek, Shi, & Yammiyavar, 2009; Hall, De Jong, & Steehouder, 2004; Smith & Yetim, 2004; Winschiers-Theophilus, 2009).

In this article, we examine how usability testing methods are applied in industry in different countries. Our research questions are as follows: Is there in the practical industrial usability work a cross-culturally standard approach to the usability test process? Does it make sense to talk about culturally specific usability test processes? In the article, we first describe related work. Then in Study 1 we report on a multisite cross-cultural field study that results in a grounded theory theoretical model of usability testing. The conclusion is that usability testing is not the same in all the locations studied. In Study 2, we aim at testing the theoretical models of usability testing developed in Study 1. This is done, 1 year later, with in-depth qualitative research interviews with key informants from the same companies as in Study 1. The results are templates for cross-cultural and culturally specific usability testing. The article ends with a general discussion of the findings from Studies 1 and 2 and their implications.

2. RELATED WORK AND BACKGROUND LITERATURE

A focus in this article is on usability evaluation methods as defined by Gray and Salzman (1998)[1] and by Hartson, Andre, and Williges (2003). In the industry,

[1]Usability evaluation methods is a broad term for analytical (without live users) and empirical (with live users) methods that the usability professional uses to evaluate the interaction of the human with the computer. The purpose is to identify aspects of this interaction that need to be improved to increase the usability of the product.

a wealth of usability and user experience testing methods are used to evaluate computer software user interfaces and other interactive products: inspection methods, workplace observation, think-aloud usability tests, and so on, with usability testing being the most commonly used (Clemmensen, 2005; Gulliksen et al., 2004). In the HCI research communities, it is not possible to publish an article on a new technology design without an accompanying empirical usability evaluation (Barkhuus & Rode, 2007). Compared to analytical evaluations with no live users (e.g., log file analysis logging clicks on websites, GOMS studies predicting the time spent on key-level interactions), empirical testing with live users (e.g., qualitative think-aloud studies revealing people's use of task-related information, quantitative studies measuring task completion time and error rates) is widely considered to be a more valid and less biased approach to usability evaluation (Barkhuus & Rode, 2007).

The assumption that we are all doing the same and getting the same results in a usability test, however, has repeatedly been shown to be wrong in studies of practical usability testing in industry. Molich, Ede, Kaasgaard, and Karyukin (2004), as part of the comparative usability evaluation (CUE) studies, asked nine independent organizations to evaluate the usability of the same website, Microsoft Hotmail. The results showed a wide difference in the selection and application of methodologies, resources applied, and problems reported. As a part of a follow-up study, called CUE-8, Kirakowski and Murphy (2009) studied measurement practices in 15 organizations that did quantitative usability evaluations of a commercial website. They found that time on task varied by a factor of four and was not comparable across the participating organizations without considerable further statistical refinement. Furthermore, the questionnaire used for measuring the satisfaction and user experience, the System Usability Scale (Brooke, 1996), had high variability and was useless for comparison across organizations.

Different explanations and solutions have been suggested for reducing the variability in the practice of usability testing and for increasing the comparability of results across testing sites. Accidental variations in the selection and application of methodology can be explained by the lack of an adequate theory to provide guidance for the usability testing practice (Boren & Ramey, 2000). Paradoxically, trying to theorize about different types of methods, such as analytical versus empirical methods, or laboratory versus field-based methods, may raise questions as to who, where, and when to apply the different usability evaluation methods (Fiotakis, Raptis, & Avouris, 2009).

Differences in the resources applied in usability testing can be explained with reference to the differences among usability teams in their view of how to provide sufficient task coverage and what representative users are (Clemmensen, 2004; Lindgaard & Chattratichart, 2007). For example, the typical participant in usability evaluations in published research articles in HCI is a male student, but recruiting a convenience sample of male students may not be sufficient even for academic research purposes (Barkhuus & Rode, 2007).

Differences in the outcome of usability testing can be due to individual differences among the evaluators, such as differences in their cognitive style, which gives a cultural bias in the test outcome (Ling & Salvendy, 2009), or differences

in the evaluators' level of expertise (Capra, 2007). Such individual differences call for more careful selection, and use, of evaluators in usability testing in order to produce valid and comparable results. In a comparison of three teams of usability professionals, who were asked to use special software support for doing comparable analysis of the results of usability tests, Vermeeren et al. (2008) found that different teams of usability professionals could not find the same sets of usability problems in the same interface. This was the case even when they were provided with software decision support for the evaluation. Vermeeren et al. concluded that the analysis process in a usability test is inherently subjective. Other HCI researchers have come to a similar conclusion about usability testing: "Usability testing now appears to be a highly variable art in which the results depend on who is testing what by which protocol with which particular subjects" (Constantine, as cited in Lindgaard & Chattratichart, 2007, p. 1417).

The influence of cultural differences on usability testing has been raised in relation to international user studies and also when introducing usability testing to new cultural contexts. In principle, it does not seem to be too difficult to perform user studies in different cultural contexts and countries. Simply extending the principles of user-centered design to the new cultural contexts could do the job. However, practical experience from international user studies shows that there are many specific issues and trade-offs that need to be made (Dray & Mrazek, 1996). Such issues have been theorized using Hofstede's (1980) national cultural dimensions. For example, Yeo (1998) describes how the existing practice derived from the West of migrating software design methods from a source culture to a target culture without any adaptation of the method may not work for usability testing in Malaysia because of power distance issues: Having a test user of lower social rank than the evaluator will result in few negative comments about the product, and having a test user with higher rank than the evaluator will result in many negative comments about the product. Ford and Kotze (2005) proposed that studies of culture's influence on usability should go beyond applying Hofstede's dimensions by building more detailed conceptual models of usability. They argued that such conceptual models are needed to guide empirical usability research in the use of the large number of variables that are necessary in cultural studies to describe variations in user characteristics, task characteristics, and characteristics of the environments, as well as strategies for controlling the variables. Ford and Kotze underscored that a more holistic view of the cultural issues that influence usability is needed. In the next paragraphs, we discuss six studies that begin to take such a view.

Clemmensen et al. (2009) applied social psychological research on cultural differences to the study of the different components of usability evaluation. Nisbett, Peng, Choi, and Norenzayan (2001), in an impressive array of empirical studies, demonstrated how people born and raised in China, Japan, or Korea tend to apply a holistic cognitive style in their thinking, whereas people born and raised in the United States tend to apply an analytical style in their thinking. An example of the difference in thinking style is that when asked to report what is in a scene, holistic thinkers mention the background, whereas analytical thinkers report the focal objects. Clemmensen et al. (2009) used this distinction to discuss what the existing

literature can say about how culture affects the different components of a usability test. They concluded that, when they have different cultural backgrounds, it is important to consider how to match the task presentation to the users' cultural background and to consider the different effects of thinking aloud on task performance in culturally different user groups, the cultural differences in nonverbal behaviour that affect usability problem detection, and finally the complexity of the overall relationship between user and evaluator.

Hall, et al. (2004) asked male students, who differed on Hofstede's (1980) national cultural dimensions by being either collectivist (African/Asian nationalities) or individualistic (Dutch nationality), to perform seven tasks on an international website for academic researchers. Observation of the students' performance showed no difference between the two cultural groups, but collectivist students self-reported significantly fewer problems compared to the individualistic students. Furthermore, the collectivist students were more reluctant than individualistic students to take on other roles (reviewer, real user) than the test user role. Although these results indicate the importance of cultural differences for usability testing, there are several factors that create doubt about the findings from the study. First, the test users were asked to think-aloud after the full set of tasks had been completed to give the nonnative English speakers time to explain themselves better. This procedure may have been interpreted differently by the two groups, for example, as a sharing activity by the collectivistic students, and as an evaluative activity by the individualistic students, and can thus have led to the observed differences in the users' role-taking behavior. Second, the procedure for usability-problem discovery was based on the test users' remarks indicating disapproval, surprise, doubt, and so forth, which may have introduced a bias in the observation of users and the count of usability problems, due to cultural differences in the nonverbal behavior between the two user groups. Finally, as the authors themselves pointed out, the African/Asian participants were foreign students who at the time of the study were living in Holland, and their answers to a survey could be biased by differences in their inclination to adjust to aspects of the Dutch majority culture. There is a risk that, in studies of usability testing with heterogeneous user groups (e.g., Law & Hvannberg, 2004), the national borders of the participant groups may be taken for granted.

Shi (2008) studied how evaluators establish a supportive relationship and communicate effectively with the users in usability testing sessions that took place in five companies in Beijing, China (the data were from the Chinese part of the field study that is reported in full here). The Chinese users focused mainly on tasks, whereas the evaluators focused on both users and tasks. Further, the Chinese users did not think-aloud actively, hence effective communication skills were needed to encourage the users to speak out. Retrospective thinking aloud and explanation were widely used in the tests. Shi concluded that the communication and relations in the Chinese usability test sessions were appropriate for formative usability evaluations but not for summative usability evaluations. She furthermore pointed out that her study was focused on verbal communication but that indirect and nonverbal communication is often assumed to be important in Asian countries. Nonverbal communication was studied by Yammiyavar, Clemmensen, and Kumar (2008), who analyzed 120 min of video from 12 thinking-aloud usability

tests in Denmark, China, and India for the nonverbal communication between the users and evaluators, and found that some types of nonverbal behaviours were used only in some of the countries. The studies indicate that culturally specific variations in the communications and relations between evaluator and test user might be important for the outcome of the usability tests in industrial practice.

Herman (1996), working as a consultant in Singapore, did a case study with usability testing of a real system developed for professional users. The users were local, and they were tested either individually or in pairs. Measures were task completion, extent of looking for help, errors, understanding of the application's semantics and concepts, effective use of the application's functions, and efficiency of interaction (e.g., use of shortcuts). Qualitative data related to the evaluation of the system were taken from questionnaires and interviews with the users. Herman found that the users expressed neutral to positive evaluations, even though their performance with the software was obviously not satisfactory. The performance appeared to depend on the social situation during the test; most effective were novice users working in pairs and the expert user in an expert–novice pair. Least effective were users being tested individually and the novice user in the expert–novice pair. However, in all cases the subjects were positive in their evaluations, including when they had performed poorly. Herman interpreted these as biased results in usability testing arising from the users' cultural background from Asia. To reduce the bias, Herman advised the use of think-aloud usability testing instead of structured interviews/questionnaires. These think-aloud usability tests should, however, be carried out only by users working in pairs, with equal social status, familiar to each other, and preferably without an observer present. Among the objections to Herman's case study is that it is unclear how the verbalization carried out by two users could be less influenced by cultural effects compared to (classical) thinking aloud by an individual user (Ericsson, 1998).

Vatrapu and Pérez-Quiñones (2006) showed the influence of culture on structured usability evaluation interviews in the United States. Users with an Indian background were not willing to talk as freely and accurately to an interviewer with a majority culture background as when they were with an interviewer with a similar, Indian background. The perceived cultural background 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 results are explained with reference to the advantage of the user and interviewer sharing the same values on Hofstede's power distance dimension. In this study, both interviewers and users apparently could speak English fluently, but the influence of language differences on the interview parts of the usability evaluation sessions is obvious. However, the issue of the effect of language use other than English in usability testing has not been studied much. X. Sun and Shi (2007) observed the effect of language in seven research laboratory sessions of think-aloud usability evaluation in Beijing. They found that speaking Chinese made the moderator assist more in the detail and encourage the users more frequently; when speaking English, the moderator had to pay more attention to the screen to understand what was going on. Similar implicit differences in language use between local-foreign pairs of evaluators-users might have contributed to the findings by Vatrapu and Pérez-Quiñones.

Hertzum et al. (2011) conducted interviews in China, Denmark, and India with stakeholders (developers, users) about their personal constructs of the software systems that they use every day. They found that the users experienced the frustrating and useful systems as similar, but the developers experienced frustrating and easy-to-use systems as similar. The Chinese participants' most characteristic use-constructs related to security, task types, training, and system issues, whereas the Danish and to some extent the Indian participants made more use of constructs traditionally associated with usability (e.g., easy-to-use, intuitive, and liked). These results pointed in the same direction as Frandsen-Thorlacius, Hornbæk, Hertzum, and Clemmensen's (2009) questionnaire study in Denmark and China, which showed that the notion of usability, its aspects, and their interrelations are not always constant across cultures, when seen from a user perspective.

3. STUDY 1

We decided to study a number of questions related to culture and usability testing: What are the current usability test practices in different countries? How do culturally different test users respond to the test instructions and test methodology? Which interfaces are more influenced by cultural diversity and current usability test practices?

3.1. Social Psychological Approach to Cultural Usability

In Study 1, we applied a social-cognitive theory of culture (Hong & Mallorie, 2004) that conceptualizes culture as a loose network of domain-specific cognitive structures (including theories and beliefs) and, furthermore, argues that an individual can hold more than one cultural meaning system, even if the systems contain conflicting cultural theories. The theory says that, depending on the accessibility, availability, and applicability of such cultural knowledge, cross-cultural differences may affect usability.

Accessible cultural knowledge is approached as meaning systems that are widely shared among members of a cultural group and frequently used in communication among members and thus become chronically accessible. In a usability test situation, where people under time pressure look for readily available and widely accepted solutions to a problem, the chronically accessible knowledge will be used and typical cultural group differences may emerge.

However, it is not only the task conditions that favor the use of chronically accessible cultural knowledge. The knowledge should also be available to the individual in the situation at hand. Individuals in many societies today are increasingly polycultural in their background. In any given situation, such an individual will carry two or more implicit cultural theories about how to perceive and act in the situation. The theory that will be applied will then tend to be the one that is most available to him or her in the specific situation. In our case, we can assume that a usability testing situation will activate certain cultural knowledge, by priming the individual test user and evaluator with culturally specific scripts,

icons, and pictures. For example, many computer applications are designed to contain culturally specific icons and pictures. These can prime evaluators and test users to apply certain types of their accessible cultural knowledge, while they complete a behavioral strategy such as a usability test.

The theory furthermore assumes that it is felt more strongly to be appropriate to apply one's cultural knowledge in situations where the other person has a similar cultural background as oneself. Hence when the users and evaluators in a usability testing session have different or similar sociocultural backgrounds, it influences the outcome of the test.

Summing up, the knowledge that test users and evaluators apply in a usability test session can be expected to vary with the knowledge that is available in the local culture, the knowledge that the usability test session materials prime, and the knowledge that is appropriate to apply in the social context of the test participants. To study these variations, we need to study the practice of usability testing at different "home grounds" across the globe. The glue that can bind substudies together is when individual researchers are present at the studies done at the other researchers' home grounds.

3.2. Research Design

In three countries (China, Denmark, and India), we approached the research questions by studying current usability testing practice in companies that test software and interactive products for the local market. We used a site-based procedure (Arcury & Quandt, 1999) for locating and recruiting qualitative study participants, to learn about each cultural setting, and to gain entry into the setting. The procedure had five steps:

1. Specify the characteristics relevant to the sample, which in our case were geographic (Denmark, India, China), sociocultural (i.e, the type of local test users who usually act as participants in usability tests), and the type of experienced evaluators (hereafter called "moderators" to emphasize their facilitating role in think-aloud usability testing). Finally, a characteristic defining our sample was related to the usability test vendor. We selected only companies that did professional user testing services of software products to the local market.
2. Generate a list of sites—in our case, the companies where the usability tests were done.
3. Estimate the composition of clientele at each site by contacting a "gatekeeper" for each site and ask for appropriate statistics for the site and help gaining an entrée. The gatekeeper(s) in our case were the manager in each company, who had the daily responsibility for running usability tests.
4. The research and gatekeeper recruit participants. In our case, this was when we and the manager agreed on when and how we were allowed to observe usability tests at the site and interview the test users and moderators.
5. Recruit individuals from sites and maintain a table indicating the characteristics of the participants in the sample to help the researcher to assess the quality of recruitment (Trost, 1986). In our case 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 any more relevant usability test vendors at the time, (b) we did achieve a reasonable amount of variation in our sample, and (c) the three geographical categories of usability test vendors were clearly independent of each other (e.g., none of the usability companies in our study directly cooperated with any of the other companies at the time of our study). The selected companies were in Copenhagen, Denmark, a local usability consultancy company with 12 employees (six tests with two different moderators); in Mumbai, India, a local branch of an international usability consultancy company with more than 200 employees (four tests with four different moderators); and in Beijing—five companies' usability groups of five to 15 employees (three local branches of international usability consultancy companies, one international IT company's in-house usability group, and one local IT company's usability group; in Beijing, in total we observed six tests with six different moderators).

In each company, we did field observation with video cameras of usability test sessions and interviews with evaluators, test users, test audience (note-takers), and usability test managers. We were three researchers: one Danish (the author), one Indian, and one Chinese. In all observations both a local and a foreign researcher were present to increase the cultural validity of the observation. When planning the observation, the usability test manager was explicitly asked to make sure that the observed tests conformed with the current practice of usability testing in that company. If it was not possible to observe a customer-paid usability test, we asked the company to redo a recently run customer-paid test. The test applications were in all cases designed for the local market.

This procedure gave us a total of 52 hr of videotaped observation in three languages across seven companies. The analysis of the interviews was done through a grounded analysis approach (Glaser & Holton, 2004), using ATLAS.ti® software (ATLAS.ti, Berlin, Germany). The goal of the analysis was not an accurate description of all data but a quest for a conceptual theory, abstract of time, place, and people done in the grounded theory tradition of Glaser and Holton (2004). The aim was to develop a grounded theory of the production of usability problems in the conduct of think-aloud usability tests.

3.3. Results

The grounded theory of the production of usability problem reports that we developed from the observations is shown in Figure 1 in a simple form. At the end of the analysis, the complete ATLAS.ti hermeneutic unit had 293 primary documents, 218 quotations, and 239 codes. In the next paragraphs we present the results on the different parts of the theory, beginning with the cultural diversity in the people taking part in the usability test sessions.

Cultural Diversity in Moderators and Test Users

The two traditional independent variables age and gender emerged from the field data as important parts of the cultural diversity among test users and

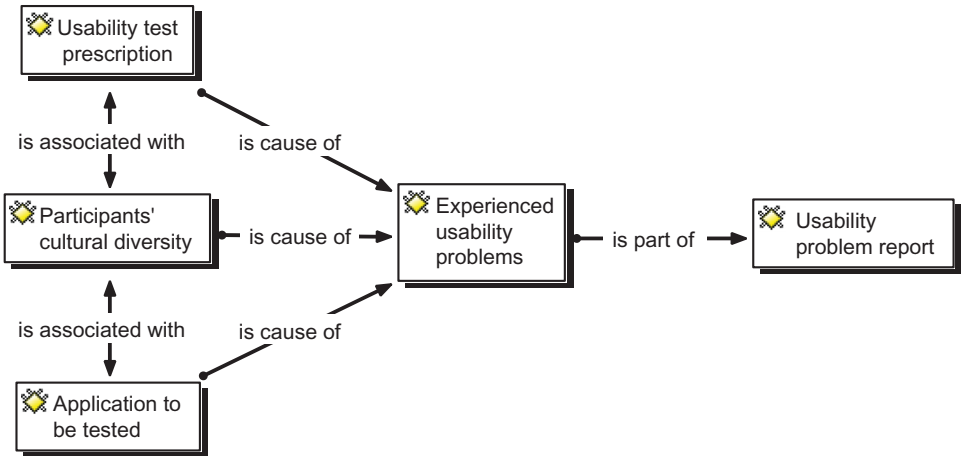

FIGURE 1 The grounded theory model of the production of a usability problem report (color figure available online).

moderators. The Indian moderators saw age and gender differences between moderator and test user as no problem with urban users, but 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 "*keen to listen.*" Specifically, a senior moderator could make a young test user with a traditional cultural background scared, which would require the moderator to "*go an extra mile to communicate*" (moderator IN 1).[2] Even in a remotely conducted usability test, in which the moderator did not see the test user, the moderators would want to know the age of their test users. Also, if the user was female and from a traditional family, a male member of her family had to be present during the preinterviews and the test, or the moderator had to be a woman: (moderator IN 1) "*It takes a woman to interview female users with traditional background.*" In the Chinese tests, age and gender differences were also emphasized: (moderator CH 1) "*. . . the way to do the test will be different . . .*" though with a different view of the role of gender differences: "*. . . if the user is male it is better if the moderator is a female. . . .*" In the Danish tests, the participants had positive feelings toward a test user who was similar to them in age, gender, and job experience—(test audience DK 1) "*He was a very good test user*"—and hostile to the extent of being offensive toward a test user who was different in age and gender—(test audience DK 1) "*She is an old . . . .*" At one point the Danish moderator could hear through the one-way screen the test audience laughing, and in the interviews he expressed his concern: (moderator DK 1) "*[that the test user could] . . . play this . . .,*" that is, he was afraid that the difference in age and gender between test audience and test user would influence the knowledge that the test user would apply and hence the outcome of the usability test sessions.

[2]From this point forward, the national shorthand notation (IN, DK, and CH) is used together with numbers (1–6) to cite the participants in the tests and interviews.

The Usability Test Prescription

Besides the cultural diversity of test participants, the grounded theory illustrated in Figure 1 suggests another main influence on the usability test: the standard textbook prescriptions for usability tests. The standard textbook prescribes that a usability test should have the following properties (Barnum, 2001): (a) specific test goals and concerns; (b) real work tasks to be done; (c) think-aloud procedures; (d) test users, moderators, observers, clients/managers, designers, in-house trainees; (e) recovery procedures; (f) moderator room and observer room; and (g) test applications and usability problem lists/reports. All of the usability tests that we observed did share this general approach to usability testing, but most exhibited interesting variations of the properties of the test.

Specific test goals and concerns. In the Danish tests, an explicit goal was to get the test users to think-aloud to allow insights into their cognitive work with the product to be tested. However, it was not clear from the beginning of the test exactly when the user was required to think-aloud. The moderator asked about the user's opinions and likings during most of a 2-hr test session. Only toward the end of the test session would he or she give the user a specific task that allowed the user to think-aloud. In the Chinese tests, it was also a goal that users should think-aloud. All the Chinese moderators said that their test was a think-aloud test. Two moderators (moderator CH 1 and CH 3) further explained that their test was a combined think-aloud and interview. However, despite their intention to run the test as a think-aloud, the Chinese moderators asked questions and the users answered throughout the whole session in all the studied companies. On occasion, the user did work on the task in silence, and then the Chinese moderator probed or asked for explanations only after the task completion. The reason for this either-ask-questions-or-do-not-interfere tactic from the moderators was given as (moderator CH 2):"*. . . there is no need to disturb the user when she takes the correct steps. . . .*" The Chinese moderator did not see it as a problem that the user did not think-aloud concurrently with the task performance, but instead did it retrospectively. In the Indian tests, it was also an explicit goal with the test that the user should be outspoken and talk about what was going on in his or her mind (which the Indian users also did). However, the Indian moderators were most concerned that usability testing should be seen as part of a user-centered design process: (moderator IN 3) "*Basically the final outcome for us has to be a tangible physical design. So for that we get more detailed information. . . .*" The design goal overruled the goal of getting the user to think-aloud to get information during the test.

Real work tasks to be done. In the Indian tests, all moderators used a detailed protocol to get the user started on the different tasks: (moderator IN 1) "*. . . through scenarios we start the task, when the user has stopped talking and he has answered most of the things, and if I have no questions to ask, the task is stopped . . .*"; (moderator IN 1) "*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 IN 2) "*I say: this is what to do, how you will 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 IN 1) "*I think it is a good task*"; (user IN 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 regarding both the start and end of the user's task: (moderator CH 2) "*. . . let's see the next task . . ." or simply ". . . after I have introduced a task, then it will begin."* The need for having a detailed protocol to guide the test was not strong: (moderator CH 3) "*. . . if you have many protocols it will scare the user, make the user very nervous, uncomfortable."* The argument was that if a moderator needs a protocol, one should ask her or him: (moderator CH 3) "*Has she done a real usability test before, or not? . . . in a real usability test, they should not have [protocols]. . . . It would scare the user. . . ."* However, another Chinese moderator was quite concerned with having and following a protocol: (moderator CH 1) "*. . . sometimes I read out the task in order to make all users get the same instructions."* In the Danish tests, there were a large number of steps in the protocol, which was partly because the client wanted a large part of the test application (a website) tested in one long test.

Think-aloud procedures. In the Danish tests, users were actively thinking aloud, for example, (user DK 6) "*. . . loan type 4,5 or 6—I do not understand any of this. . . ."* The Danish moderators did not give many reminders to think-aloud. Instead, they gave affirmative reminders, for example, (moderator DK 2) "*no," "uhmm," "yes," "okay."* However, the users tended to mix thinking aloud with giving their opinion about the test product: (user DK 4) "*. . . this overview is quite good . . ."*; they also suggested design improvements: (user DK 4)"*I would have written. . . ."* In response to this kind of comment from the users, the Danish moderators probed for more information, for example, (moderator DK 2) "*Are there other things you think of when you see this homepage?"* Some of the moderators' probes were so general that they appeared to be more suitable for an interview: (moderator DK 1) "*What do you think when you see this homepage?"* When a user had a silent period, for example, (user DK 6) "*[the user is trying something with the calculator to solve a task and does it silently and finally says]. . . . I think it does do it wrong . . .,"* the moderator would probe for possible intentions or actions: (moderator DK 2) "*What will you do here now?"* The think-aloud procedures in the Danish usability tests appeared to be quite relaxed and more like interview procedures.

In the Indian tests, users were actively thinking aloud and speaking out: (user IN 1 repeatedly) "*. . . now I will do . . .,"* though they were sometimes thinking in phrases: (user IN 1) "*I think in today's age. . . ."* When necessary, the moderator used reminders to help the user to think-aloud. Reminders could have many forms, as in Table 1.

All the Indian moderators used their hands and arms to make lively gestures to support their speech. The moderators were trained in body language: (manager IN) "*. . . moderators are trained about 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 Indian tests, several ways of probing for more information came out:

Table 1: Reminders to Think Aloud, From Usability Tests in an Indian Company

Reminders and Probes
• Please think aloud
• What are you looking for in this page?
• What is happening now?
• Whatever you like or dislike or you think, you can say
• Can you say what you are finding
• Talk to me, what are you looking at?
• Keep talking
• Could you give me more information about what are you doing?
• What you are expecting?

1. When the moderator reacts to the user's initiative, for example, (moderator IN 1) "*You just said that right,*" (user IN 1) "*Yes, yes, I mean this thing. . . .*"
2. When the moderator gives direction to the user by asking, (moderator IN 3) "*Did you notice that?*" and then points to something on the screen, and the user answers, "*No, that was my mistake, I didn't.*"
3. When the moderator actively wants to help the users (moderator IN 1): "*Usually in the first task I always ask such questions in order to let the user know what his job is, and later I don't do that because by then the user should know what he should do.*"
4. When the moderator actively wants to dig deeper into the user's thinking aloud, for example, when a user says, "*I am not happy with the description . . .*" (user IN 4), and the moderator asks, "*Why is that?*" (moderator IN 4).

The distinctions between asking questions, reminding to think-aloud, and probing are not necessarily maintained in the current usability test practice in India.

In the Chinese tests, it appeared to be necessary to use a kind of retrospective reminder. In one company, the tasks were very simple search tasks that required the user to find articles on a website. This was a very easy task for the user, so she completed all the tasks very quickly (between 2 and 5 s per task), and she did not say anything during task performance. Then the moderator gave a retrospective reminder: (moderator CH 4) "*You have finished the task, what did you think just now?*" In another company, when the user tried to edit text (input some words), and was silent for maybe more than a 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, (user CH 3) "*I have never done it [the task] before. . . .*" In a third company, the user did not think-aloud at the beginning of the first task, and then the moderator reminded the user twice: (moderator CH 2) "*. . . you can talk [about] what you are doing . . . which input method are you using now . . .*" and only then did the user begin to think-aloud: (user CH 2) " *. . . Oh, I've found it, I should fix it, and click the button. . . .*" Paradoxically, almost all the users in the interviews explicitly stated that they focused on doing the assigned task and were not concerned about whether the moderator understood them or not, but many times the users, after having been silent for some time, wanted to explain what they had

been doing to the moderator. In sum, the think-alouds in the Chinese tests were mostly retrospective think-alouds or explanations (Ericsson & Simon, 1993).

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, and head of the usability company that did the test, was observing the whole test. He wrote the usability report as a Microsoft PowerPoint show for each user, concurrently with the usability test, and in front of the test audience. Part of the reason that he gave for this behavior 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 IN) "*The relationship between the test user and moderator should be that of a teacher and a student, the user should be the teacher and the moderator the student–master and apprentice, actually. . . .*" The user in the Indian tests was supposed to take the role of a design critic who evaluates the website, and teaches the moderator about the usability problems. In the Chinese tests, the focus was on the test user–computer dialogue: In two companies, the user did not seem to consider the moderator at all, and almost never looked at the moderator, not even when the user said something, for example, when he or she answered questions from moderator, he or she would continue looking at the computer screen. Almost all users focused on the task. They thought the moderator was just a person who gave them instructions and facilitated the test, and they considered him "*just a little more than if I was answering a questionnaire from an anonymous sender . . .*" (user CH 2). 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 the people who participate in usability tests are well educated, so they will know the purpose of the test, and that they are not the subject, but the product is; and (c) the users were explicitly told that the test was about the product, not themselves.

Recovery procedures. In the Indian tests, a concern was that the user should not be stuck unnecessarily on a task. The moderator helped if necessary: (user IN 4) "*The two times I got stuck he helped me. . . .*" Similar behavior was identified in the tests that were observed in the Danish and Chinese companies.

Moderator room and observer room. In the Danish tests, the observer room accommodated up to 10 clients, observers, trainees, researchers, and more, who through a one-way screen could watch the moderator and the test user in the moderator room. A video of the test user, a PC screen capture, and the sound were played on monitors in the observer room. A note-taker in the observer room took notes, freeing the moderator in the moderator room from taking notes during the session. In the Indian and Chinese tests, the arrangements were very similar to the Danish tests, with a separate note-taker in the observer room. In China, the test in one company was different from the test in the other companies, because there was no dedicated note-taker. In this case, the moderator was also note-taker and wrote the report afterward. Another variation was found in another company, where

the moderator was in the observer room looking through the one-way screen and interacting only through a button-operated telecom with the test user, who was alone in the moderator 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 PowerPoint, compiling the different users' responses, and structuring the report according to the order in which the different areas of the test application were being tested. The moderator was, however, consulted before the note-taker finalized the report to the external client. In two of the Chinese companies the moderators used a template to write and present the test report to the in-house clients. In the case of external clients, a full report would be made. In the Indian company, usually the note-taker would use a predefined Microsoft Excel template, with the test application areas from the test protocol listed in a row, and the users in columns. The users' reactions, performance scores, and satisfaction measures could be entered directly in the spreadsheet and then later used to produce the report. The Indian usability report would usually consist of a 50- to 80-page text document and a PowerPoint presentation.

The Test Application

Besides the cultural diversity of test participants, and the textbook prescriptions for usability tests, the grounded theory illustrated in Figure 1 suggests a third main influence on usability test: the product or application to be tested in the usability test.

Across all the tests, it appeared that the test users' work with the test application was most efficient when they had previously been fixed on a certain, typical use of the functions of the test application. This observation extends findings from basic research in psychology and anthropology, which shows that subjects living in a technologically sparse community become fixed on the design function of the object immediately after even a brief demonstration of the object's function (German & Barrett, 2006). It also supports the findings from Rau, Choong, and Salvendy (2004) that novice computer users from different cultural groups are not necessarily comparable but can best be seen as novices in comparison to expert computer users with a similar cultural background. If you are a novice user in a culture that surrounds you with computers, this is not comparable with being a novice user in a culture with very few computers.

We observed that the test users in usability tests will often be urban, modern, young, with higher education, fluent in English, and with much computer experience. In the Indian tests, the test application was a newspaper 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 of using this kind of website. In the Danish tests, the test application was a website for an Internet and telecommunications provider. The users had higher education, 25 to 55 years of age, and end-user to medium to professional expert use-experience with the test

application. They spoke their local language but also spoke English. In the Chinese tests, the test application in one company (CH 4) was an e-learning public school website and the test user was female, young, not fluent in English, with experience of similar applications. In another company (CH 3), the test application was a search engine website, and the user was male, with higher education, not fluent in English, and with much experience in the test application but not in the tested new functions. In the third company (CH 1) the test application was web-based chat software, and in the fourth company (CH 2) it was mobile phone interfaces (pen vs. key); in both these companies the users were young, female, highly educated, fluent in English, and with extensive knowledge of similar applications. In the fifth company (CH 5) the test applications were a web-based work flow tool, with a male user in his 30s, highly educated, an in-house employee, with experience from similar systems, English speaking, and a mobile phone service provider's customer website with a test user who was young and had experience of similar applications. Obviously, in the current practice of usability tests there is awareness of recruiting users who culturally meet the test "applications' affordances" (Norman, 1988). One observation that we made, however, is that when new software is being tested, users are not always sufficiently fixed on the design function of the software, so that even when the user knows the kind of application, he or she may have little clue about the intended use of the specific functions being tested, whereas other users of the same application may actually have a clear idea about the functions being tested. This variation in the recruited test users' ability to meet the affordances of the test applications was probably a cause of variations in what we saw of the experienced usability problems. Finally, some test applications are made for other user groups and require the recruitment of test users with other backgrounds like elderly/children and/or rural, traditional, strongly religious, low education, local language only, and no computer experience. Such test users may have no or very few computers in their daily environment, and hence they are very open, flexible, and not fixed on what computers can do.

Experienced Usability Problems and the Usability Report

The cultural diversity in moderators and tests users, the usability test practice, and the affordances of the test application together contribute toward what is experienced as usability problems and towards what is considered the final usability report that is given to the client (see Figure 1). This multicausal model of experienced usability problems partly contradicts existing usability problem theory, which says that the list of usability problems reflects only the properties of the product being tested (Preece, Rogers, & Sharp, 2007). However, cultural diversity in the test users and moderators (Law & Hvannberg, 2004) and variations in the test setup (Kjeldskov, Skov, Als, & Hoegh, 2004) have been shown to influence the detection of usability problems. Our study supports this. In the Indian and Chinese tests, usability problems were experienced in interactions between test users and moderators, that is, as codiscovery episodes. In the Indian tests, users, when asked, came up with one or two major problems that they believed were there during the tests. For example, (user IN 2) found that he needed help from the moderator to solve the task: "*[The moderator] did give me indications . . . like where would you go to search for flights. . . .*" In the Chinese tests, the user directly

suggested design changes. For example, a moderator (moderator CH 2) asked the user, "*. . . why did you not find it just now . . .,*" then the user (user CH 2) said "*Oh, I didn't notice this part . . . ,*" then the moderator asked, "*. . . that means they should not put this in this part?*" and the user replied, "*. . . yes, they should put it in a flash or something. . . .*" In the Danish tests, the users were quite confident that they could identify usability problems by themselves; they did not refer to the interactions with the test moderator but instead focused on their own needs as users. These differences in the production of the usability problem list could mean the moderator–user relation extends beyond the session and into the postsession period of producing the usability problem list.

3.4. Discussion

Compared to previous studies of culture and usability testing (e.g., Vatrapu & Pérez-Quiñones, 2006), this study is distinctive in its multisite approach. A theoretical model of the production of usability problem reports in seven usability test vendors across three countries was constructed through a grounded theory analysis, which included systematic use of researchers with different cultural backgrounds, and checking concepts with the participants to increase validity. This model covers cultural diversity in test users and moderators/evaluators, variations in the textbook conduct of usability tests, and assessment of user-technology fit, in a holistic conceptual framework for appreciating nuances in the outcome of usability tests in different regions of the world.

The use of a grounded theory approach to compile the observations from different field study settings helped conceptualize usability testing based on field observations of real usability testing sessions in three countries. The conceptualization was discussed with the people that we observed doing the usability testing, which lends some validity to the grounded theory approach. However, a weakness in grounded theory is that it allows for many different relations between concepts to be considered in one theoretical model. For example, in our grounded theory we had "is a cause of," "is a part of," "is associated with," and more relations. The sheer number of different theoretical interpretations associated with having several different relations in one theoretical model made the grounded theory model a challenge to interpret. We decided therefore to test the grounded theory in a more narrowly designed follow-up field study. This Study 2 should focus in depth on only one theoretical relation between concepts, which should make the results easier to interpret and compare across field study settings, and in addition provide specific advice on how to do usability tests in different cultural settings.

4. STUDY 2[3]

In Study 2 we wanted to examine the grounded theory model of usability testing that was developed in Study 1. To make the interpretation of the results easier, we decided to focus only on the relation: "is part of." We would test systematically how the different elements of the theoretical model from the first study

[3]Parts of Study 2, in an earlier version, were published in Katre, Orngreen, Yammiyavar, and Clemmensen (2010).

were "a part of" usability testing practice in different countries. Our research questions were as follows: What is part of a standard usability test in India? What is part of a standard usability test in Denmark? What is part of a standard usability test in China? What is part of a standard usability test across all three countries?

4.1. Method

This study was done as a follow-up study 1 year after Study 1. The three companies from Study 1 selected for this follow-up Study 2 were (a) a Mumbai-based company with more than 200 usability and user-centered design specialists that is an Indian branch of an international usability consulting company, (b) a Copenhagen-based usability vendor with 12 employees, and (c) a Beijing-based branch of a major telecommunications international company with an in-house group of usability specialists.

Our initial ethnographic records and grounded theory model from Study 1 was the basis for a taxonomic and a paradigm analysis (Spradley, 1979) of what is a part of a usability test in the selected companies. In each of the three companies, we did 2 days of ethnographic follow-up interviewing with our key informant from the year before. In all three cases the key informant was a usability evaluator with senior responsibility. We followed a classical ethnographical interview procedure suggested by Spradley (1979):

1. Create a network/set of codes related to the code "Usability test" by an is-a-part-of relation.
2. Print a code hierarchy (a specific procedure in the ATLAS.ti software used in the analysis).
3. Ask the informant questions about each term (code) in the hierarchy: name, other of same kind, difference from others, and so on.
4. Do it for one subdomain at a time.
5. Enter all the responses in the code hierarchy.
6. Go back, change the network of codes accordingly.
7. Iterate the process if necessary.

In the Day 1 interview, we created, adjusted and verified the taxonomy by asking the informant structural and contrast questions, such as (Spradley, 1979):

- Is <X> a term (code) you would use?
- Would most people here at <this company> usually use this <X> term?
- Is <Y> a part of <X>? Are there different parts of X? What other parts of <X> are there?
- Do you see any differences between and and <X.1> and <X.2> and <X.3>? (etc.)

The <X> term could, for example, be on the highest level of the taxonomy "usability test" or on a lower level, for example, "inform participant."

In the Day 2 interview, we created the paradigm by this procedure:

1. Place the first level of the taxonomy in a column in a worksheet.
2. Inventory all other codes related to "usability test" by other relations than is-a-part-of relations and place them as the top-row in the worksheet.
3. Prepare contrast questions such as, "Is moderating a usability test dependent on the test user's age or gender?"
4. Conduct an interview with the informant to elicit the needed data.

The final step was to use the analysis to discover general cultural themes. This was done by careful analysis and comparison of the interview data.

To give an example of the procedure, in the interviews with the Indian informant the final taxonomy showed that 182 concepts are part of a standard usability test in the studied company. Of these, 23 were main concepts that had up to three sublevels.

Figure 2 illustrates how a part of a standard usability test in the company in Mumbai is to make the participant (the test user) comfortable. The purpose of this is to get the user to "open up." The "open-up" activity varies in duration depending on the user. It is also used to get users to think out loud in the proper way.

The full paradigm for the usability test was then made by setting up a matrix. The rows of the matrix were the different parts of the usability test taken from the main elements in the taxonomy, and the columns were the context factors. For example, how exactly important parts of a usability test in Mumbai, such as greeting, compensation, and thanks to the participant, will be carried out depends in the Mumbai company on the context factor of the user's gender: "*. . . I cannot shake hands with a lady. . . .*" This and other findings are presented in more detail in the next sections of this article.

Comparing the taxonomic and paradigm analysis across the three countries was done in two steps. First we described the common parts (taxonomy analysis) and common context (paradigm analysis) of a usability test across the three countries. The result was a cross-cultural paradigm for a usability test with cross-cultural parts (the parts mentioned in all countries) and cross-cultural context (the contexts mentioned in all countries). Second, we described for each country, in addition to the cross-cultural elements, the specific parts, and specific contexts for

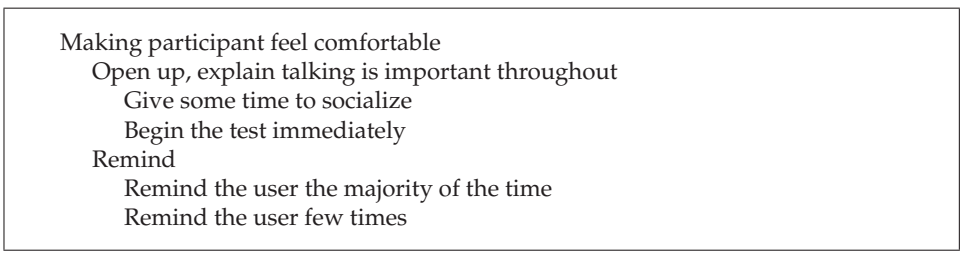

FIGURE 2 Excerpt from a taxonomy of usability test parts that shows the part: "Making the participant feel comfortable," with subparts.

a usability test in each country. The results were three country-specific paradigms for a usability test process.

4.2. Cross-Cultural Findings

This section presents the parts and context factors for a usability test that are common across the three countries. There are four main parts of a usability test, each of which has a number of subparts (Table 2).

The influence of context factors on one or more parts of the usability test is seen in the paradigm in Table 3. Cross-culturally there are eight context factors that in some way or another influence the four parts of a usability test.

In the following, the interview data on the four parts and the eight contextual factors are discussed individually.

First, already during the recruitment of test participants it is important to consider how to introduce the to-be-tested technology solution to the future test user. The Indian informant pointed out that "*the client may define a target audience that requires that . . . [we consider issues of] . . . computer literacy or application awareness.*" Furthermore, the think-aloud technique may sometimes require what the Danish informant described as "*smooth talkers,*" and she said that it may be important during recruitment to tell the future test users that there will be an observation room with people from the client company.

Second, the users' diversity influences the parts of a usability test. The instruction and task part is related to age; the Indian informant described the situation like this: "*Introducing that to older people requires a bit more explanation sometimes, this is how it is going to work, they are not very tech savvy as such. . . .*" Age does also play a role for verbalization, as the Chinese informant said: "*Older persons need more encouragement,*" and for expected task performance when observing the user: "*Young or very old people are not expected to be able to solve the same tasks as a standard adult*" (Danish informant). Of interest, despite the commonsense nature of the statements, there are few studies that deal with usability across the life span (see,

Table 2: The Interview-Based, Cross-Cultural Taxonomy for a Think-Aloud Usability Test

Main Part of Usability Test	*Subpart of Usability Test*
Instruction and tasks	Introduce user to think-aloud Introduce user to technology to be tested Introduce user to test task
Verbalization	Probe for specific information Remind the user to think aloud Communicate with the user
Reading the user	Observation room, one-way mirror Video of user and screen Expectations of user's task performance
Overall user–evaluator relationship	Explain user not tested, design tested

Table 3: The Cross-Cultural Paradigm for a Usability Test

	Context for Usability Test							
Usability Test Parts	*Recruitment of Participants*	*User's Age*	*User's Personality*	*User's Cultural Background*	*Formative Evaluation*	*The Final Report*	*Consolidate the Data*	*Design Recommendation*
Instruction and tasks	Yes	Yes	Yes	No	Yes	Yes	No	No
Verbalization	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reading the user	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Overall user-evaluator relationship	No	No	No	No	Yes	No	No	No

Note. Yes = that part of the usability test is influenced by the context factor; No = it is not influenced by the context factor.

e.g., Brouwer-Janse, 1995). The few findings on specific age groups like the elderly show that they are just as technologically savvy as the rest of the population (Rousseau & Rogers, 1998). We have not been able to find studies that compare usability testing with different age groups, which could disprove or confirm our informants' views.

Third, the user's personality is a contextual factor that is relevant for giving "Instructions and Tasks," "Verbalization," and the "Overall User–Evaluator Relationship." For Instructions and Tasks, "*You would like to give the same [instructions], but in a different language [to users with a different personality]*" (Danish informant). For Verbalization, "*It depends on the extrovert or introvert nature of a person, if he or she feels comfortable with verbalizing . . .*" (Danish informant). Both the Danish and Chinese informant felt that when and how to encourage and stop a test user from speaking depends on the test user's personality. Finally, "*If he is an introvert, your body language and everything will move towards making him comfortable . . .*" (Indian informant). These statements can be compared to experiences from other fields that apply evaluation methods. For example, in the field of administering psychological personality tests, the tester is expected to adapt the communication with the test taker according to his or her personality. However, so far few studies of usability evaluation have included personality, and some studies briefly state that they consider personality unimportant (e.g., De Angeli, Sutcliffe, & Hartmann, 2006). One exception is a study of interactive television prototypes that were designed in color and shape to show different degrees of extrovertness and evaluated in reference to the users' measured personality traits (Arvid, 2002).

Fourth, the user's cultural background influences the Verbalization and Reading the User parts. Usability evaluators may experience that "*in Singapore the users are more shy than in the US. . . . [We do not] give as many reminders [to think-aloud] in the US as in Singapore and India*" (Indian informant). A difference in cultural background is something the evaluator can use: "*. . . you may use the difference, you can take the role of a stranger entering from outside*" or something to be learned: "*. . . culture can be a professional qualification that you do not have*" (Danish informant). Cultural background can also be viewed as a matter of differences in education: "*. . . [cultural background matters] if we're talking about education, people with low education need more encouragement*" (Chinese informant). There seem to be three aspects of the concept of cultural background: national/ethnic culture (Indian informant), professional culture (Danish informant), and educational culture (Chinese informant). Despite the multiple meanings of the concept, we kept the cultural background as a context in the cross-cultural paradigm for pragmatic reasons (see Marcus, 2006).

Fifth, the test methodology acts as a context for all the parts of a usability test in the sense that all four parts of a usability test are fixed in a summative evaluation/beta test, but in formative evaluation or design evaluation the properties of the test parts vary. The variation is related to how to instruct the user in thinking aloud, how to tell the user about the paper prototype, if there are any real test tasks or only interview questions, how much the user should think-aloud, and how much qualitative data are needed. In a formative test, the expectations to task performance are plastic: (Chinese informant) "*. . . yes, if they say something outside the expectation, you should reverse the expectation, if all the results go way off . . . you*

can stop and reverse it." Finally, in a formative test, the necessity of making the user comfortable varies. The existence of such variation in formative usability testing, which is not the case in summative usability testing (where all parts of the test are performed in a fixed way), indicates a need for investigating more how this variation in the context of a usability test influences the test procedures. This aligns with recent suggestions to limit research on criteria for evaluating usability testing methods to formative usability testing (Hartson et al., 2003).

Sixth, the final three contextual factors in Table 3 all relate to considerations about how to communicate the results. What to write in the final report, how to consolidate the data, and how to present the design recommendations all influence the parts of a usability test. In some cases, the instructions and tasks are written in the final report. The final report usually also contains information about the verbalization: (Danish informant) "*. . . usually we write if we had to lead [the test users] a lot, if they were helped or not, if they acted spontaneously or not . . .*"; (Chinese informant) "*. . . if probing happened [we write it] . . . especially if there is a common response from that kind of user, this kind of user needs more encouragement. . . .*" Also information about how the user was read/understood is entered into the final report: (Indian informant) "*. . . if the client has certain expectations we show them the graphics [on user performance].*" Expectations about typical ways of consolidating the data (e.g., that reports typically provide excerpts of user verbalization) and ways of giving design recommendations (e.g., results are often presented using PowerPoint with wireframe examples) both influence how the verbalization occurs and how the user is read. This mutual influence, from consideration about how to communicate the results, to the main parts of a usability test (and the reverse), indicates that usability testing is carried out with the philosophy of iterative testing. It is congruent with recommendations from research on communicating the results of usability tests to designers, which says that evaluators should be explicit about the data behind their claims, but not overwhelm the designers with information, and rather involve them in a learning process (Nørgaard & Høegh, 2008).

4.3. A Cross-Cultural Template for the Usability Test Process

The information from the interviews analyzed earlier indicates that a usability test is a complex affair. A simple count of the cells in Table 3 gives $4 \times 8 = 32$ possible aspects for consideration by usability vendors when they carry out a usability test. The interviews and the previous analysis indicate, however, that not all of these 32 aspects are equally important. Figure 3 depicts graphically (the area painted black) the 21 important aspects of the four parts of a usability test (the 21 aspects with a "yes" in Table 3). The depiction can be regarded as a paradigm or a *template* for a cross-cultural usability test process.

The template for a cross-cultural usability test process that is shown in Figure 3 can be interpreted as follows. If a practitioner (or a researcher) is going to perform a cross-cultural usability test, at a minimum he or she has to consider the aspects that are colored black in Figure 3. For example, he or she should ask the question, "How will the think-aloud verbalization that I require from the user support the design recommendations that I will give?" The gain from having a graphic

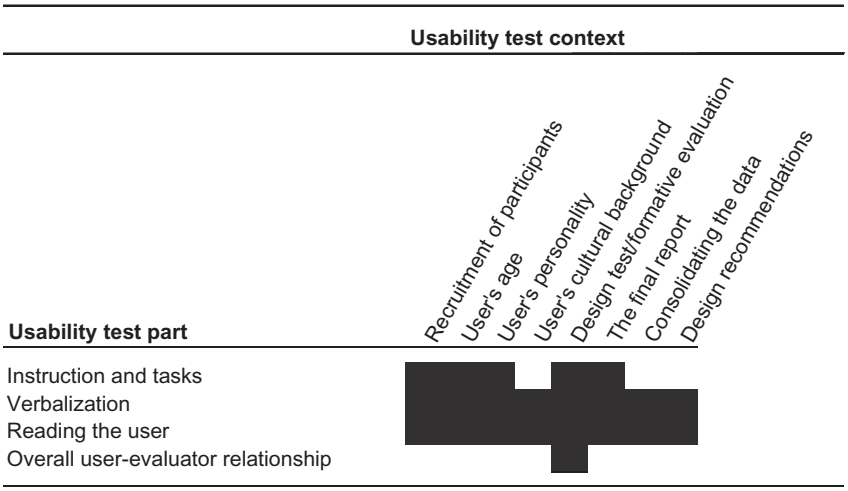

FIGURE 3 The cross-cultural template for a usability test.
Note. A graphic view of the information presented in Table 3.

depiction of the cross-cultural usability test process is that it can be used as a must-have checklist in the preparation of tests. Furthermore, a representation of the type illustrated in Figure 3 is useful when looking for variations in how the usability test should be carried out in different cultures and countries.

4.4. Culturally Specific Parts of a Usability Test

In each country, the informant mentioned parts and contexts that were not mentioned by the informants in the other countries. Although this should not be generalized to suggest that all usability tests within a given country are performed in a similar way (and different from the way that tests are performed in other countries), the interview data indicate three distinctly different ways of performing a usability test. Further analysis of the data suggested that it makes sense to distinguish between an "evaluator-centered," a "user-centered," and a "client-centered" approach to the usability testing process. These three approaches are described in the following sections.

Beijing: An "Evaluator-Centered" Approach to the Usability Test

The Beijing informant told us about parts and contexts of a usability test that we could not recognize from our studies in the other countries (see Figure 4). First, the "Choice of usability testing method" could vary for intranet, web search portals, and government web pages "*. . . a few methods like expert review can be used . . .*" but also for cross-cultural contexts. For example, "Choice of usability testing method" was about fitting the method to the user's personality "*. . . this kind of user is more talkative, maybe more useful in some kind of tests . . .*" or to the user's cultural background "*. . . [according] to not only country or religion, but education, different*

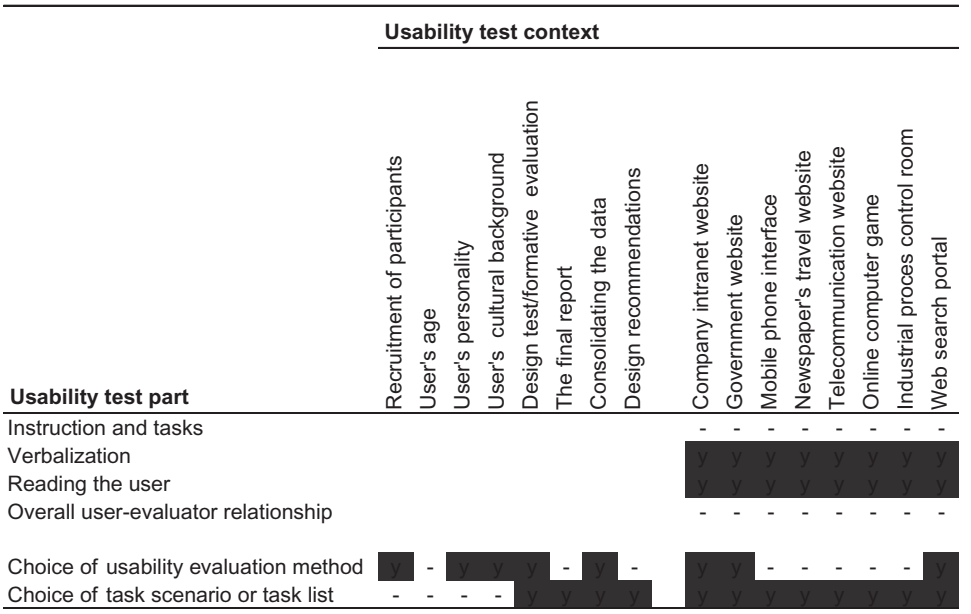

FIGURE 4 The "evaluator-centered" template for a usability test in Beijing.

people can give different feedback, like the IT person gives totally different feedback from the general user . . ." and to the user's lifestyle and family background "*. . . products like sports, select the natural observation method. . . .*" Second, the "Choice of task scenario or task list" both varied with technology solutions as a matter of fitting topic with methodology, and would accordingly be written in the final report as part of data consolidation and design recommendations.

The technology solutions to be tested were important contexts for the standard parts of a usability test. "Verbalization" was related to all kinds of technology solution "*. . . but only through the methodology, sometimes you need more qualitative data. . . .*" "Reading the user" was related to who would be the observers of that technology "*. . . if it's a hot topic, the marketing people should hear directly from the users. . . . [If intranet or other internet technology] the designer has no need to hear it directly from the user. . . . [If it's mobile phone interfaces] usually the designer will join in . . . .*" When asked about who chose the methodology and why the methodology was most important, the informant explained that the evaluator chooses the methodology in each case of a usability test and that this was possible because the evaluators were usually highly qualified human factors professionals. Together with the findings from Study 1 this suggested, for the template shown in Figure 4, the label "evaluator-centered" usability test process.

Mumbai: A "User-Centered" Approach to the Usability Test

In Mumbai, in addition to the cross-cultural usability test parts the informant explained about 10 other parts of a usability test (see Figure 5).

The task scenario part of a usability test is influenced by whether summative or formative tests are done, and it is usually described in the final report. The task

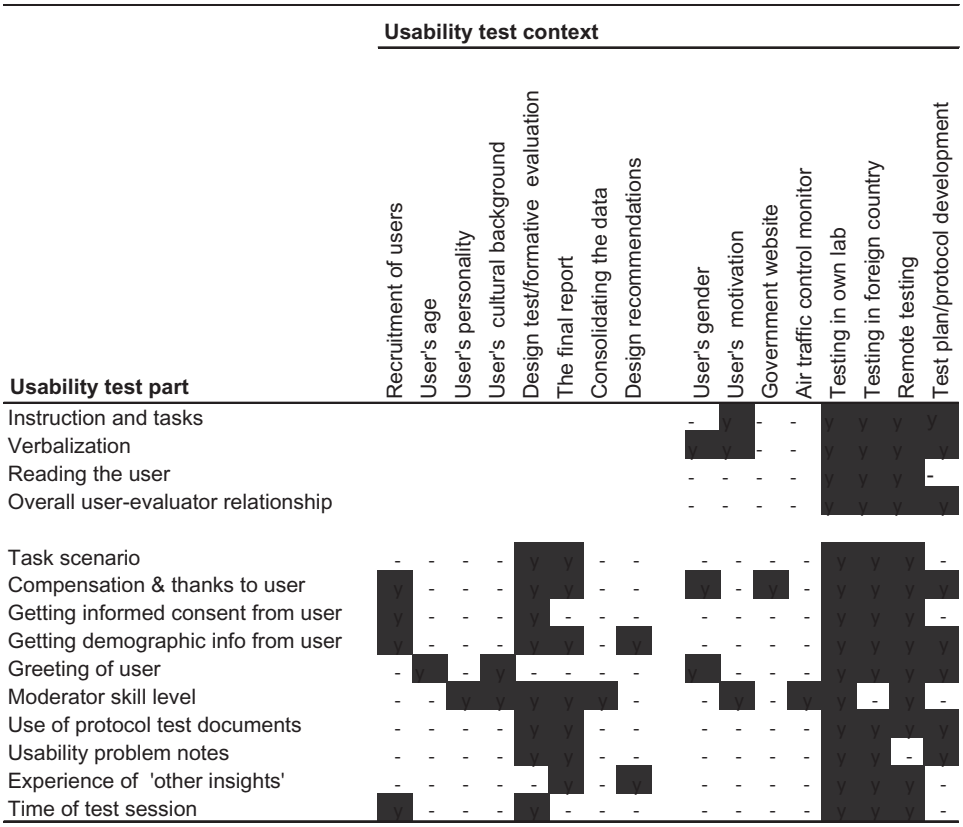

FIGURE 5 The "user-centered" template for a usability test in Mumbai.

scenario is also influenced by where geographically the test takes place, which was a context-factor that was mentioned only by the Indian informant.

There are four distinct parts of a usability test that are related to dealing with the user as a person. The conduct of these parts varies in different contexts. For example, when recruiting the user "*. . . we think about what kind of compensation is to be given . . . to a certain extent we ask them if they are willing to come here . . . nothing in writing but we have the consent . . . we are asking permission [from the family] . . .,*" and when greeting the user "*. . . in the US, for example, even if he is an older or younger person, the greeting will be the same. Here the elderly will feel good if I bow. . . .*" Specific contexts for the Mumbai usability test include gender: "*. . . I cannot shake hands with a lady . . .*"; government websites, "*. . . in the US, if your client is a government agency, you cannot give them compensation . . . [actually] I don't know if this is the case in India. . . .*"

The moderator skill level is related to the user's personality: "*. . . if [the user] is an introvert, you [the moderator] may need to . . .*"; the user's cultural background: "*. . . You may need to know the nuances of the culture and understanding of that . . .*"; test methodology: "*. . . for formative tests it would be good to have an experienced moderator . . . a summative test is fairly straightforward. . . .*"; writing the final report: "*. . . it*

is part of the contract that we have experienced moderators . . . also an experienced moderator is involved in creating the final report . . . very closely . . . various models here, he could be writing it, he could be overseeing [parts of] the final report . . ."; consolidating the data: "*. . . the note-taker in consolidation with the moderator, usually the moderator is more senior than the note-taker. . . .*" For the moderator's skill level there are also specific contexts such as the user's motivation: "*. . . in the sense that you have to realize that when a person is not motivated, he is probably not giving you the real feedback . . . the moderator has to realize that . . . and he has to do some twists . . .*"; the kind of technology solution: "*. . . if it is a complex application we would need experienced moderators . . .*"; and remote testing: "*. . . it is good to have someone who has done some remote testing, because the technology issues . . . making a phone, call, the supporting things, you should be aware of the things that can go wrong during the test, . . . whether it be phone line, internet connection, web example, accent. . . .*"

The use of test documents such as formal test protocols and notes about usability problems is also an important part of the usability test in the Mumbai company. These have to be visible in the final report and data consolidation: "*. . . if you have not been able to conduct all the tasks as per the protocol, you have missed out something. . . . You do mention what data will be captured . . . in remote testing you will not be able to capture body language and facial expressions. . . .*" Getting other insights about the user's interaction with the technology solution is also a standard part of a usability test, in particular in formative tests: "*[The data are] much richer here than in summative tests, things can strike you here. . . .*" The duration of the test session is a standard part of the test that is considered during recruitment of test users: "*I just mention to the user that it will take one hour . . . .*"

Eight Mumbai company-specific context factors were mentioned by the informant (see the right-most eight columns of Figure 5). These eight context factors influenced both the 10 culture-specific parts of the test and the four cross-cultural parts of the usability tests. For example, "Verbalization" depends on the user's motivation: "*. . . [if low] very strongly, a lot more probing would be required, giving more reminders, also correlation to assists . . . .*" and is considered during test protocol development, when interacting with the client: "*. . . Sometimes the clients are saying that we really need you to get more information . . . then we identify where more probing is required . . . .*" The "Overall user–evaluator relationship" is related to the user's motivation and considered during test protocol development: "*You have to make extra efforts if that person is not motivated. . . . One of the reasons for using [specific kinds of] scenarios is to make the user comfortable. . . .*"

The heavy focus on the test parts related to greeting and informing the user and the test context factors' significant influence on the overall relationship between evaluator and test user inspired us to label the template illustrated in Figure 5 a "user-centered" approach to the usability test process.

Copenhagen: A "Client-Centered" Approach to the Usability Test

The findings from Copenhagen are shown in Figure 6. Besides the moderator's experience with clients, the Danish informant mentions using "clickable prototypes" as an important part of usability testing that is always there, no matter what the different contexts are.

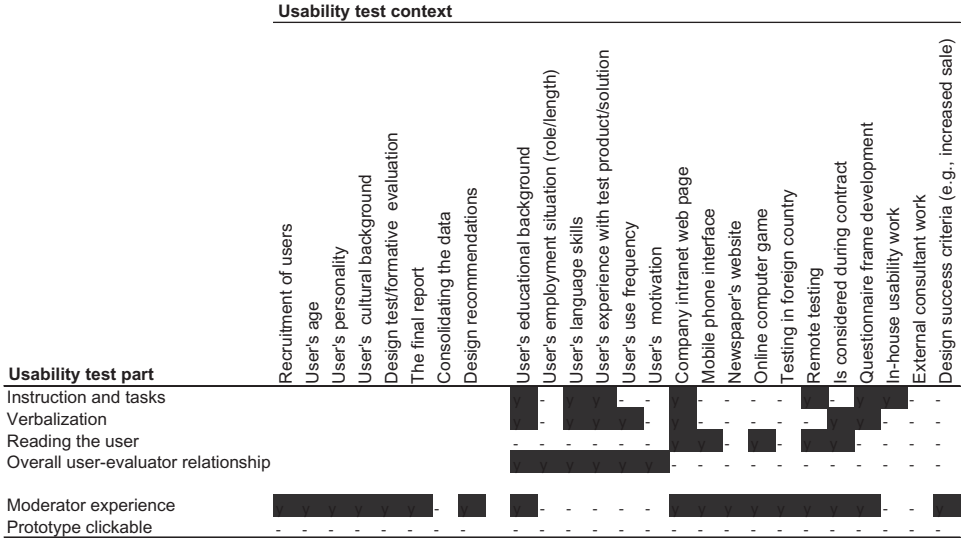

FIGURE 6 The "client-centered" template for a usability test in Copenhagen.

Among the 17 different contextual influences on a usability test mentioned by the Danish informant is the user's gender: "*. . . females over 40 years are less talkative . . .*" ("Verbalization"), and "*. . . usually we select both male and female users because their context of use can be different*" ("Reading the user"); educational background: "*. . . you would like to express yourself slightly differently . . .*" ("Instructions and tasks")," *. . . higher education gives more verbalization and better verbalization, people with lower education have a tendency to apologize for not being able to do the test correctly . . .*" ("Verbalization"), and different expectations of task performance ("Reading the user"); the user's employment situation (role and length) "*. . . there will be different tasks for managers and employees, there will be some things that employees cannot answer, it is not part of their job . . .*" ("Reading the user"); the user's language skills: "*the problem can be that you cannot be sure of understanding what they are saying . . .*" ("Instructions and tasks") and "*. . . if their English is so bad that they cannot read the task instructions . . .*" ("Reading the user"); the user's experience with the technology to be tested, use frequency and motivation: "*. . . if it is a super user, I must say 'tell me about your knowledge' . . .*" ("Verbalization"), and expectation of task performance ("Reading the user"); intranet web pages: "*. . . in a test of the intranet you have to make a point to the user that he or she is anonymous, and try to explain to them that in the report it will not be written who they are . . . In such a test you often speak about use and frequency of use. . . .The boss might not like to hear that . . .*" ("Instructions and tasks"), "*. . . [in intranet tests] you are the stranger coming from outside and have to try to figure out how much you know the concept and tools that are there . . . can you be part of that organizational culture . . .*" ("Verbalization"), and "*. . . there is no observation room because the test [of an intranet] will be done in the field, not in the lab . . .*" ("Reading the user"); mobile technology: "*sometimes*

there is no observation room, because we move around [during a test of mobile units] in town, in traffic . . ." ("Reading the user"); online computer gaming: "*. . . you may go to the user's home to watch them gaming in their environment . . .*" ("Reading the user"); remote usability testing: "*. . . you have to explain a bit technically how the test will proceed, how you as a user get access, and that at the end of the test you will be disconnected . . .*" ("Instructions and tasks") and "*Usually there is no observation room, and if there is one, the observer has also logged on with a separate screen . . .*" ("Reading the user"); contract negotiation: "*. . . we write about the method in the contract . . .*" ("Verbalization"), "*. . . we always write [in the contract] that we have an observation room and customers like that we are open and trust them, that they can come and learn about their users by observing them . . .*" ("Reading the user"); questionnaire guide: "*. . . then there will be a script that tells us how to talk to these users. . .*" ("Verbalization"); and in-house usability work: "*I clearly explain that the technology solution is not my design, so I will not be offended [by the user's critique]*" ("Instructions and tasks").

Finally, the context of success criteria (increased sale, interested community users, number of users seeing the advice given, etc.) is relevant for a part of the usability test that only the Danish informant mentioned, namely, the moderator"s experience of working with clients: "*. . . there is not sufficient focus on that [client's success criteria] if you have taken a route through usability in your education—instead you need to have experience working with clients, you need to know what makes your clients pay attention to this and to be persuaded that this is necessary . . . It is important for your design recommendations that you reflect on what the business goals are for the use of this technology solution. . . .*" This focus on business goals, together with the view of the legal contract with the client as an important context for carrying out the usability test, suggested the use of the label "client-centered" approach to the usability test process.

4.5. Discussion

The ethnographic interviews with the taxonomic and paradigm analysis indicate that a typical usability test across countries has some clear similarities, with four main parts and eight important contexts to consider when doing the test.

In addition to the four main parts, in each country there are specific parts of a usability test and specific contexts for a usability test, which are not found everywhere. In Mumbai, most parts of the usability test were related not to the interactive application that was tested but to differences in user characteristics, test preparation, method, and location. In Copenhagen, considerations about the client's needs were always part of a usability test. In Beijing, a specific part was the evaluator's choice of evaluation method.

To cite a methodology insight from this study: Compared to the current popular grounded theory approach, the strength of the ethnographic approach is that the terms revealed are the informant's own terms. Thus tacit knowledge developed in the company was revealed by our study. This is, on the other hand, also the major limitation of this follow-up study. Only one informant in one company was interviewed in each country. However, these interviews were each 2 × 3 hr of in-depth

interviews held across 2 days and based on previous extensive field observations. Another possible weakness is that the informant in each country may have attempted to do his or her best to conform to some international standard, or what the informant believed he or she had learned during formal education, that is, the informant may have violated the interviewer's need for a "nonanalytic informant" (Spradley, 1979, p. 52). Then again, we actually wanted an analytic informant who could give us the official version of how to do a standard usability test in that company. A third methodological challenge, which has been met with some success in this study, is to compare ethnographic studies from three sites distributed globally.

5. GENERAL DISCUSSION

The research questions were as follows: Is there cross-culturally in industry practice a standard approach to the usability test process? And/or does it make sense to talk about culturally specific usability test processes? In response to these questions, a theoretical model of the production of usability problems in seven usability test vendors across three countries was constructed, based on field studies. This model includes cultural diversity in test users and moderators, variations in the conduct of the usability test, and assessment of the user-technology fit, in a holistic conceptual framework for explaining the outcome of usability tests in different regions of the world. Follow-up ethnographic interviewing 1 year later confirmed that a typical usability test across countries has some clear similarities, with four main parts and eight important contexts to consider when doing the usability test. A graphic view of a template for cross-cultural usability testing was presented (Figure 3). Furthermore, in each country we found specific parts and specific contexts for a usability test. Three templates were presented in a graphic form: The "evaluator-centered" usability test in Beijing (Figure 4), the "user-centered" usability test in Mumbai (Figure 5), and the "client-centered" usability test in Copenhagen (Figure 6).

5.1. Cross-Cultural Templates for Usability Testing

The template for cross-cultural usability testing, with four main parts and eight important contexts to consider when doing the usability test, suggests that there is agreement across countries as to what senior usability professionals consider are parts of a usability test. The identification of a cross-cultural template for usability testing disagrees with the variability in all parts of the usability testing process that has been suggested by the CUE studies (see, e.g., Kirakowski & Murphy, 2009; Molich et al., 2004).

Some parts in the cross-cultural template may vary with the context of the test, and some may not. Here the study supports the finding from related work (Clemmensen et al., 2009; Hall et al., 2004; Hertzum, 2010; Hertzum et al., 2011; Shi, 2008; Vatrapu & Pérez-Quiñones, 2006) that it is important to consider how the cultural context influences the user's willingness to speak out loud from different perspectives, and how to read the user's expressions of usability and user experience problems. It is in contrast to (Herman, 1996), who recommends the use

of standard verbal protocols also in Asia. Furthermore, despite that much recent research indicates that we need to take cross-cultural differences into account (e.g., Choi et al., 2006), the cross-cultural template shows that the lesson from literature studies that some users prefer scenario-based task presentations and others prefer task lists (Clemmensen et al., 2009) has not been accepted as a general rule of usability testing practice everywhere. As the cross-cultural template in Figure 3 shows, there is no agreement across countries that users' cultural background should be considered when giving instructions and tasks.

5.2. Culturally Specific Templates for Usability Testing

We suggest that a large portion of the "considerable inherent subjectivity" in usability testing proposed by Vermeeren et al. (2008, p. 369) may be due to systematic differences in the parts and context of usability testing in different countries. Thus, most of the variation in usability testing processes is perhaps not subjective in the sense of being random or arbitrary. Rather, it may follow common practices that have emerged against the background of the history and culture of that region of the world. We have identified a systematic variation in the practice of usability testing in different countries and described this variation in three culturally specific templates for usability testing.

The "evaluator-centered" usability test in Beijing had two specific parts (choice of usability evaluation methods, and choice of tasks scenario or task list) and eight specific context factors that were all technology specific. The Beijing template was the only one that included in the usability testing that the evaluator should choose the evaluation methods. Interviews revealed that the historic background for this was that many of the usability professionals in Beijing, when usability testing became widespread, were recruited from engineering psychology research institutions and preferred to choose their own methods. This indicates the importance of more research into the history of the usability profession in different countries in order to understand how usability work is carried out. The other culturally specific part in the Beijing template, choosing between a task scenario or task list for task presentation, is shared with the template from Mumbai. The two templates thus confirm the literature-based finding that Asian users prefer scenario-based task presentations (Clemmensen et al., 2009).

The "user-centered" usability test template from Mumbai had 10 specific parts and eight contexts to consider. Most of these are related to dealing with users and test locations not familiar to the evaluators, such as when testing in a foreign country or testing with rural users. Much of the practical advice concerning relevant trade-offs given by, for example, Dray and Mrazek (1996) is thus covered by this template. However, why exactly in Mumbai the users' level of motivation is related to the moderators' skill level remains to be explained. Here the research into the role of communication and relations in usability testing (e.g., Shi, 2008) may be of help. Yet more research is needed to provide a theoretical account of how gender plays a role in greeting and compensating the user, and what the difference is between doing tests in one's own versus a foreign country versus remote testing.

The "client-centered" usability test in Copenhagen had two specific parts and 17 specific context factors. One of the parts, "use of clickable prototypes," was not context dependent at all. The other specific part, "moderator experience," varied with nearly all the context factors. This seemed to be similar to the focus on the evaluator in the Beijing template, but the interviews revealed that in the Copenhagen template, the kind of relevant experience was, in particular, experience with contract negotiation and handling the payment of clients. This kind of usability professional experience or evaluator skill has rarely been studied in the context of culture, but see Mayhew and Bias (2005) for a discussion of relevant factors that may enter such a discussion.

6. CONCLUSION

There are two points in conclusion:

1. Across countries there is agreement as to what senior usability professionals consider to be part of a usability test. This agreement can be described as a template with parts that are constants and parts that vary depending on the context of the usability test.
2. For each of the studied locations for usability testing—Copenhagen, Mumbai, Beijing—a culturally specific template for usability testing can be created that shows what specific parts to add to the test and what additional contexts to consider for each part of the usability test.

6.1. Implications

It is important to realize in comparative usability evaluation that the goal can be to create value for users and evaluators as well as for clients. This is important because of the systematic variation in how usability testing is carried out in practice in different countries. Realizing the systematic nature of the variation, and using the templates presented in section 4, can help us understand that not all variation in usability testing is due to unwanted and irremovable subjectivity but can be explained and dealt with. Developing templates for cross-cultural and culturally specific approaches to usability testing for more countries can help the emergence and development of the local usability profession in the different countries. This is crucial for adaptation of the concept of usability testing to new cultural contexts.

REFERENCES

Arcury, T., A, & Quandt, S. A. (1999). Participant recruitment for qualitative research: A site-based approach to community research in complex societies. *Human Organization*, *58*, 128–133.

Arvid, K. (2002). Personality preferences in graphical interface design. In *NordiChi 2002* (pp. 217–218). Aarhus, Denmark: ACM.

Barber, W., & Badre, A. (1998). Culturability: The merging of culture and usability. In *4th Conference on Human Factors and the Web*. Retrieved from http://research.microsoft.com/en-us/um/people/marycz/hfweb98/barber/

Barkhuus, L., & Rode, J. (2007). From mice to men—24 years of evaluation in CHI. In *CHI 2007 (Alt-CHI)*. Retrieved from http://www.itu.dk/~barkhuus/barkhuus-altchi.pdf

Barnum, C. M. (2001). *Usability testing and research*. Needham Heights, MA: Allyn & Bacon.

Beu, A., Honold, P., & Yuan, X. (2000). How to build up an infrastructure for intercultural usability engineering. *International Journal of Human–Computer Interaction*, *12*, 347–358.

Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*, 261–278.

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.

Brouwer-Janse, M. D. (1995). From our past to our future: User interfaces over the lifespan (panel session). In *CHI 1995* (pp. 187–188). Denver, CO: ACM.

Capra, M. G. (2007). Comparing usability problem identification and description by practitioners and students. In *Human Factors and Ergonomics Society Annual Meeting 51* (Vol. 51, pp. 474–478). Baltimore, MD: Human Factors and Ergonomics Society.

Choi, B., Lee, I., & Kim, J. (2006). Culturability in mobile data services: A qualitative study of the relationship between cultural characteristics and user-experience attributes. *International Journal of Human—Computer Interaction*, *20*, 171–203.

Clemmensen, T. (2004). Four approaches to user modelling—A qualitative research interview study of HCI professionals' practice. *Interacting with Computers*, *16*, 799–829.

Clemmensen, T. (2005). Community knowledge in an emerging online professional community: The case of Sigchi.dk. *Knowledge and Process Management*, *11*(2), 1–10.

Clemmensen, T., Hertzum, M., Hornbaek, K., Shi, Q., & Yammiyavar, P. (2009). Cultural cognition in usability evaluation. *Interacting with Computers*, *21*, 212–220.

Clemmensen, T., & Plocher, T. (2007). The Cultural Usability (CULTUSAB) Project: Studies of cultural models in psychological usability evaluation methods. In N. Aykin (Ed.), *Usability and internationalization. HCI and culture* (Vol. 4559, pp. 274–280). Beijing, China: Springer Berlin/Heidelberg.

Constantine, L. (2003). CHI 2003 Feature: Testing . . . 1 2 3 4 5 . . . Testing. . . . In *CHI 2003*. Retrieved from http://usabilitynews.com/news/article1058.asp

De Angeli, A., Sutcliffe, A., & Hartmann, J. (2006). Interaction, usability and aesthetics: What influences users' preferences? *DIS 2006* (pp. 271–280). University Park, PA: ACM.

Dray, S., & Mrazek, D. (1996). A day in the life. Studying context across cultures. In E. del Galdo, & J. Nielsen (Eds.), *International user interfaces* (pp. 242–256). New York, NY: Wiley.

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*, 178–186.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal reports as data*. Cambridge MA: MIT Press.

Fiotakis, G., Raptis, D., & Avouris, N. (2009). Considering cost in usability evaluation of mobile applications: Who, where and when. In *Interact 2009* (pp. 231–234). New York, NY: Springer.

Ford, G., & Kotze, P. (2005). Researching culture and usability—A conceptual model of usability. *People and Computers*, *19*, 317–333.

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. In U. Boston (Ed.), *CHI 2009* (pp. 41–50). New York, NY: ACM.

German, T. P., & Barrett, H. C. (2006). Functional fixedness in a technologically sparse culture. *Psychological Science*, *16*(1), 1–5.

Glaser, B. G., & Holton, J. (2004). Remodeling grounded theory. *Forum Qualitative Sozialforschung / Forum: Qualitative Social reseach [Online Journal], 5*(2), 4. Retrieved from http://www.qualitative-research.net/fqs-texte/2-0472-04glaser-e.htm

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, *13*, 203–261.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., & Herulf, L. . (2004). Making a difference - A survey of the usability profession in Sweden. In *NordiChi 2004* (pp. 207–215). Tampere, Finland: ACM.

Hall, M., De Jong, M., & Steehouder, M. (2004). Cultural differences and usability evaluation: Individualistic and collectivistic participants compared. *Technical Communication*, *51*, 489.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria For evaluating usability evaluation methods. *International Journal of Human–Computer Interaction*, *15*(1), 145–181.

Herman, L. (1996). Towards effective usability evaluation in Asia: Cross-cultural differences. In *Proceedings of the 6th Australian Conference on Computer–Human Interaction (OZCHI'96)* (pp. 135–136). Sydney, Australia. Washington D.C.: IEEE Computer Society.

Hertzum, M. (2010). Images of usability. *International Journal of Human–Computer Interaction*, *26*, 567–600.

Hertzum, M., Clemmensen, T., Hornbæk, K., Kumar, J., Shi, Q., & Yammiyavar, P. (in press). Personal usability constructs: How people construe usability across nationalities and stakeholder groups. *International Journal of Human–Computer Interaction*.

Hofstede, G. (1980). *Culture's consequence: Comparing values, behaviours, institutions and organizations across nations*. London, UK: Sage.

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.

Jagne, J., & Smith-Atakan, A. S. G. (2006). Cross-cultural interface design strategy. *Universal Access in the Information Society*, *5*, 299–305.

Katre, D., Orngreen, R., Yammiyavar, P., & Clemmensen, T. (Eds.). (2010). *Human work interaction design: Usability in social, cultural and organizational contexts*. New York: Springer-Verlag.

Kirakowski, J., & Murphy, R. (2009). A comparison of current approaches to usability measurement. In *UPA 2009 Workshop: Comparative Usability Task Measurement (CUE–8)*, Portland, OR.

Kjeldskov, J., Skov, M. B., Als, B. S., & Hoegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Mobile Human-Computer Interaction—Mobilehci 2004, Proceedings* (Vol. 3160, pp. 61–73). Berlin, Germany: Springer.

Law, E. L.-C., & Hvannberg, E. T. (2004). Analysis of combinatorial user effects in international usability tests. In *CHI 2004* (pp. 9–16). Vienna, Austria: ACM.

Lindgaard, G., & Chattratichart, J. (2007). Usability testing: What have we overlooked? *CHI 2007* (pp. 1415–1424). San Jose, CA: ACM.

Ling, C., & Salvendy, G. (2009). Effect of evaluators' cognitive style on heuristic evaluation: Field dependent and field independent evaluators. *International Journal of Human–Computer Studies*, *67*, 382–393.

Marcus, A. (2006). Culture: Wanted? Alive or dead. *Journal of Usability Studies*, *1*(2), 62–63.

Marcus, A., & Gould, E. W. (2000). Crosscurrents: Cultural dimensions and global web user-interface design. *Interactions*, *7*(4), 32–46.

Mayhew, D. J., & Bias, R. G. (2005). Cost-justifying usability engineering for cross-cultural user interface design. In N. Aykin (Ed.), *Usability and internationalization of information technology* (pp. 213–252). Mahwah, NJ: Erlbaum.

Molich, R., Ede, M. R., Kaasgaard, k., & Karyukin, B. (2004). Comparative usability evaluation. *Behavior and Information Technology*, *23*(1), 65–74.

Nisbett, R. E., Peng, K. P., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*, 291–310.

Norman, D. A. (1988). *The design of everyday things*. New York: Basic Books.

Nørgaard, M., & Høegh, R. T. (2008). Evaluating usability: using models of argumentation to improve persuasiveness of usability feedback. In *DIS 2008* (pp. 212–221). Cape Town, South Africa: ACM.

Preece, J., Rogers, Y., & Sharp, H. (2007). *Interaction design: Beyond human–computer interaction* (2nd ed.). New York, NY: Wiley.

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.

Rauterberg, M. (2006). From personal to cultural computing: How to assess a cultural experience. In *Proceedings of the 4th Usability Day (9. Juni 2006)*. Applied University Vorarlberg, Dornbirn, Austria: Pabst Science.

Rousseau, G. K., & Rogers, W. A. (1998). Computer usage patterns of university faculty members across the life span. *Computers in Human Behavior*, *14*, 417–428.

Shi, Q. (2008). A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests *NordiChi 2008* (pp. 344–352). Lund, Sweden: ACM.

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), 1–5

Spradley, J. P. (1979). *The ethnographic interview*. Fort Worth, TX: Harcourt Brace Jovanovich.

Sun, H. (2004). *Expanding the scope of localization: A cultural usability perspective on mobile text messaging use in American and Chinese contexts*. Unpublished doctoral dissertation, Rensselaer Polytechnic Institute, New York.

Sun, X., & Shi, Q. (2007). Language issues in cross cultural usability testing: a pilot study in China. In N. Aykin (Ed.), *Usability and internationalization. Global and local user interfaces* (Vol. 4560, pp. 274–284). Beijing, China: Springer Berlin/Heidelberg.

Trost, J. E. (1986). Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative Sociology*, *9*(1), 54–57.

Vatrapu, R., & Pérez-Quiñones, M. A. (2006). Culture and usability evaluation: The effects of culture in structured interviews. *Journal of Usability Studies*, *1*, 156–170.

Vermeeren, A. P. O. S., Attema, J., Akar, E., de Ridder, H., von Doorn, A. J., Erbug, C., . . . Maguire, M. C. (2008). Usability problem reports for comparative studies: Consistency and inspectability. *Human–Computer Interaction*, *23*, 329–380.

Winschiers-Theophilus, H. (2009). The art of cross-cultural design for usability. *Lecture Notes in Computer Science*, *5614*, 665–671.

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. (1998). Cultural effects in usability assessment. In *CHI 1998* (pp. 74–75). Los Angeles, CA: ACM.

Zakaria, N., Stanton, J. M., & Sarkar-Barney, S. T. M. (2003). Designing and implementing culturally-sensitive IT applications. *Information Technology & People*, *16*(1), 49–75.