Cross cultural usability testing

The relationship between evaluator and test user

Clemmensen, Torkil; Goyal, Shivam

Working Paper

Cross cultural usability testing
- the relationship between evaluator and test user

By

Thorkil Clemmensen & Shivam Goyal

No. 06 - 2005

Institut for Informatik

Handelshøjskolen
i København

Howitzvej 60
2000 Frederiksberg

Tlf.: 3815 2400
Fax: 3815 2401
http://www.inf.cbs.dk

Department of Informatics

Copenhagen
Business School

Howitzvej 60
DK-2000 Frederiksberg
Denmark

Tel.: +45 3815 2400
Fax: +45 3815 2401
http://www.inf.cbs.dk

Cross cultural usability testing
- the relationship between evaluator and test user

Torkil Clemmensen,
Department of Informatics
Copenhagen Business School
Denmark
tc.inf@cbs.dk

Shivam Goyal
Department of Design
Indian Institute of Technology
Guwahati, India
shivam@iitg.ernet.in

ABSTRACT
In this paper, we present the results of a pilot study in Denmark of cross cultural effects on Think Aloud usability testing. We provide an overview of previous research on cross cultural usability evaluation with a special focus on the relationship between the evaluator and the test user. This relation was studied in an experiment with usability testing of a localized clipart application in which eight participants from Denmark and India formed pairs of evaluator-test user. The test users were asked to think aloud and the evaluators' role were to facilitate the test users thinking aloud and hereby identify usability problems with the clipart application. Data on the evaluators' and test users' behaviour were recorded and analyzed by coding and summarizing statistics on these behavioural events. The results show that Think Aloud Usability Test of a localized application is most effectively performed, in terms of number of think aloud events and number of usability problems found, when both the evaluators and the test users are local. These results are however limited to the Danish context and need to be investigated in other cultural settings.

Author Keywords
Cross Cultural Usability Testing, International Usability Testing.

ACM Classification Keywords
H5.2. User interfaces: Evaluation/methodology; H.1.2 User/Machine Systems: Human Factors; I.3.6 Methodology and Techniques: Ergonomics.

INTRODUCTION
Having test users to Think Aloud during usability testing is generally thought to be an effective and successful technique (Schneiderman et al. 2004). According to Nielsen (cited in Boren et al. 2000), Think Aloud usability testing is the single most valuable usability engineering method for evaluating the usability of user interfaces. However, the descriptions in the Usability Literature of how this method is followed in the industry do not conform to the theoretical basis of this method (Boren et al. 2000). The theoretical basis of the method is described in 'Protocol Analysis: Verbal Report as Data' (Ericsson et al. 1993; Ericsson et al. 1998). A certain relaxation from theoretical rigor when we apply a scientific method to practical purposes is what we may expect.

What is less understood and accepted, however, are the significant differences in how people from different cultures respond to directions and test methodologies. With the advent of globalisation and IT revolution, we can no longer overlook the aspect of culture in the design of user interfaces and products (Russo et al. 1993). Taking cultural issues into account has now become one of the key factors for the success or failure of a global product. Despite this fact, however, we don't have any kind of formal method which guides us to evaluate a product to a certain standard while keeping in the sensitivity to cultural issues in different countries. International Usability Testing just involves a usability expert from one country and a local evaluator in the target country (Nielsen 1990). The purpose with the research presented here is to develop and/or adjust existing recommendations for the conduct of Think Aloud usability test so that this test can be used with a cultural diverse user population, be outsourced to usability companies working off-shore and/or be used in global companies employing usability evaluators with diverse cultural backgrounds.

Previous studies on cross cultural usability evaluation have shown that culture broadly affects the usability evaluation assessment processes. In one sense, we try with the ongoing research presented in this paper to look into the broad issues raised by (Smith et al. 2004): "How do we avoid cultural bias in requirements elicitation and usability data collection?" and "What user based evaluation methods address cultural diversity in both the moderator and user?" Specifically, in this paper we are interested in where to start an investigation of the assumption that the usability evaluator almost needs to belong to the target culture to (a) completely understand how people will respond to the Think Aloud directions and test methodology and (b) understand what is the effective way to obtain test users' usability feedback, without actually disguising the usability problems.

In the rest of the paper we first briefly review the literature on culture and cultural models and the literature on general usability test methods in an international perspective, and then see both in relation to the think aloud literature. Then we report on an experiment with usability testing of a localized clipart application in which eight participants from Denmark and India formed pairs of evaluator-test user. The paper ends with a discussion of why the local evaluator –local test user relation seems to be the most effective in terms of the generation of think aloud events and usability problems.

Culture, cultural models and interfaces

When we consider culture, we may begin with opposite ideal types, e.g. the 'universalist' vs. the 'relativist' views (Shweder et al. 1991). In the 'universalist' camp, we find (Bourges-Waldegg et al. 1998), who concludes that culturally determined usability problems in interfaces are due to the understandings of the representations and that those meanings lay in the culture specific contexts. The current internationalization and localisation process that is used in the design of universally usable HCI systems is in their view not quite appropriate because:

- Its overdependence on guidelines
- The difficulty of determining the user from the present cultural grounds, as culture are dynamic and keep interacting with each other.
- Its tendency to build stereotypes which later become design rules
- Its treatment of different cultures with one specific language that doesn't take in account cultural heterogeneity.

The majority of cultural breakdowns thus occur in user-tool interaction and user-task interaction, because of test users' lack of understanding of the representations of tools and tasks. Cultural differences are basically representational differences. Cultural factors (such as religion, government, language, art, marriage, sense of humour etc.) are present in every culture, but the ways they are represented vary from culture to culture, in the view of (Bourges-Waldegg et al. 1998).

In the 'relativist' camp, we find (Vöhringer-Kuhnt 2002), who studied the attitude of Usability professionals from various countries, towards usability components viz. Efficiency, Effectiveness, and Satisfaction. He found that usability professionals from different countries have specific inclinations towards one of these components; for them any usability study primarily concerns that specific component. Vöhringer-Kuhnt suggested these inclinations are due to the cultural influences. Vöhringer-Kuhnt also studied how cultural specific variables influences users' beliefs about software effectiveness; users' perception about software efficiency; users' attitude of satisfaction; and how these culturally specific variables correlate to each other and affect the overall attitude towards the usability of the product. There are no significant correlation between cultural specific variables and components of usability and also no significant correlation between the attitude towards components of usability and attitude towards overall usability. However, the Cultural specific variables significantly correlate with the attitude towards overall usability.

Besides those that can be categorized as 'universalist' or 'relativist' interpretations (or cultural representation vs. culturalization perspectives (Dormann et al. 2002)), there have been other attempts to model and measure cross cultural differences between various cultures. One frequently cited model is Geert Hofstede's cultural model. The study of (Marcus et al. 2000) on Hofstede's culture dimensions and how they affect user-interface design (their study was based on web sites from various countries) showed that we need to change our current practices in UI design and we have to develop new tools that can incorporate processes and methods for developing multiple versions of a website in a cost effective manner.

Previous work in cross cultural usability and usability evaluation

The reliability of Usability tests has been questioned from the evaluators' perspective by (Hertzum et al. 2001; Hertzum et al. 2000). According to them, the usability testing technique suffers from a major defect called the Evaluator Effect: the total number of usability problems found will depend upon the knowledge and experience of the evaluator and on the number of evaluators. (Kessner et al. 2001) puts a question mark on the reliability of usability testing by asking how many teams of

evaluators that are needed to find all the usability problems and how much agreement they will reach and on what basis. This 'evaluator effect' may be associated with cultural factors.

Herman (Herman 1996) studied the effect of culture on objective and subjective usability evaluation. She suggested that cultural factors significantly affect the correlation between subjective and objective evaluation. She concluded that subjective evaluation should be augmented by objective tests, and furthermore that subjective evaluation techniques should be used with caution as the reliability of these is susceptible to cultural effects. Testing subjects individually should be avoided, as little information may be retrieved. She found that verbal protocol techniques are most effective when the tests are conducted using subject pairs who are familiar with one another.

Herman's findings are similar to those found in the studies of Yeo (Yeo 1998; Yeo 2001). His aim was to identify, examine and reduce the effect of cultural factors that influence usability testing. Initial results showed that an important possible cultural factor is power distance: a test user who was of higher rank than the experimenter gave more negative comments about the product than the one who was of lower rank than the experimenter. To get honest results from usability testing, the experimenter should therefore be of the same rank or of lower rank than the test subjects.

In another paper, (Yeo 2001) argued that the existing practice (derived from the West) of migrating software from a source culture to a target culture may not be appropriate. According to him this works in the design and implementation phase, but NOT in the usability evaluation phase. He employed three Usability Assessment Techniques (UATs): Thinking-aloud Technique (for objective evaluation), System Usability Scale (SUS) and Interview (for subjective evaluation). The results of the Usability evaluations were found to be inconsistent. These inconsistencies arose due to the factors: 'computer experience', 'power distance (large)' and 'collectivist (as opposed to individual)' nature of the test users. According to Yeo, the cause of these inconsistencies was the participants' reluctance to provide critical negative comments. They were reluctant because they wanted to 'preserve the face' of the designer and because they showed respect for hierarchy. The results of Yeo's studies imply that to obtain data in high PDI (Power Distance Index) and collectivistic countries, UATs which collect objective (like TA) data should be used. If one wants to obtain data using subjective measures, then it is important to use participants who are experienced in tools similar to the product being evaluated, and who are familiar to the experimenter.

(Vatrapu 2001) found that in international usability enquiry with structured interviews, the culture of the interviewer had an effect on the number of usability problems found, on the number of suggestions made, and on number of positive and negative comments made. Vatrapu & Pérez-Quiñones (Vatrapu et al. 2004) found that those participants who are from the same culture as that of the interviewer (India) brought more usability problems than participants who are interviewed by the interviewer who was not of the same culture (in this case Anglo-American). In a similar line of research, (Miller et al. 1994) asked: Do language and cultural differences between staff and participants negate the outcome of usability tests? Are foreign nationals good representatives of users in their home country? Finally, as a warning, we will emphasize that (Khaslavsky 1998) raised the important issue that we now only study west and east and their cultural differences but in west also the culture of USA is far different from European countries.

(Nielsen 1990) discussed the issues for making localised interfaces, and came up with the conclusion that localization is not just mere translation of text, it's more than that. To make usable user interfaces the localized interface should be made using the usability engineering methods similar to those used in the development of original user interface. Nielsen suggests that to conduct international usability tests: travel to the target country yourself or conduct the test remotely or hire a local usability consultant to run the test for you.

Thinking aloud and culture

In their study of usability testing in practice, (Boren et al. 2000) questioned the appropriateness of the classical cognitive account of what is means to Think Aloud given by (Ericsson et al. 1993) in their book: 'Protocol Analysis: Verbal Report as data' . This classical model says that during the Usability test session there should be very little interaction between a test user and a evaluator. After a task begins, the only kind of interaction should be to ask the user to keep thinking aloud.

Boren and Ramey argued, however, that a evaluator doing a usability test should not act as a passive listener, because a speaker, here a test user, cannot ignore listeners and expects response, agreement, sympathy, execution etc. Accordingly in the interaction between the evaluator and test subject, the evaluator should pretend that test subject knows the interface and the evaluator want to learn from him or her. Boren and Ramey said that to establish such a communication, we have to use the technique called Speech Genre Model, which is based on the following key points:

- The subject of the test is the interface, not the user.
- The test user is the expert of the work that the interface is for, and the evaluator is the Learner.
- While the test user is verbalizing, he should be acknowledged, time to time, by the evaluator, using speech tokens.

- These tokens should not take the speaker-ship from the test user.

(Anderson 2004) pointed out that the purpose with the interaction is to make user keep thinking aloud. For that he suggests non-directive, open ended questions such as "What are you thinking?", "Is that what you expected?", "What has just happened?" and we should avoid questions such as "Are you confused?" or "Were you trying to copy the file?" as they act as very directive and run the risk of interpreting behaviour. According to (Ericsson et al. 1993) the degree of interaction depends greatly on the kind of test, e.g. a quantitative test would be much less interactive than a qualitative one. However, as Nielsen et al. (Nielsen et al. 2002) pointed out, not even a well executed Thinking Aloud test is sufficient to get access to what a test user is thinking. We think faster than we talk and TA interferes with the task performance. Thought processes are much more complex than the verbalized output.

An alternative usability test has been proposed by (Chavan 2004), who calls her Usability Testing method the 'Bollywood Method'. This method is especially tailored for test users from India. In this approach a user is provided with a situation in which he/she has to complete the task in order to achieve something (depending upon the situation). This is done in order to motivate test user from India to get a quality feedback from them.

THE EXPERIMENT

Based on our study of previous research we conclude that culture significantly affects the usability evaluation process and that the present usability test theory and test methodologies do not take cultural issues into account and therefore need to be improved to make them more culturally sensitive.

There are a number of issues in cross cultural usability testing that previous research has only touched upon. The 'Evaluator Effect' (Hertzum et al. 2001; Hertzum et al. 2000), i.e. the variability in evaluators' detection and selection of usability problems, is already an issue in Usability Testing research, so it should come as no surprise that we may also have a 'Cross cultural evaluator effect'. There are many issues that revolve around evaluators: What role plays the culture of the evaluator in relation to the culture of the test user in a Cross Cultural Usability test? As cultures keeps changing and evolving from one generation to the other, and some cultures faster than others at a given time in history, what role does the age of the evaluator in relation to the age of the test user in a fast changing culture play? What function does the gender of the evaluator in relation to the gender of the test user play in a usability test in different cultures?

Another issue in Cross Cultural Usability Testing is the 'User Effect' (Law et al. 2004), which is the capacity of user to detect usability problems in a particular usability test, depending on their knowledge and experience. This issue also has many layers which are undiscovered. Is it perhaps very important to use groups of test users in one culture, while in another culture it will work fine with individual test users? And if groups of test users are needed, then how does the test user pairing, e.g. novice-novice, expert-novice, expert-expert, have effect on the results of a cross cultural usability test? (Herman 1996) How will these groupings change the outcome of the test as we move from east to west? Is the evaluator aware of the fact that easterners are more comfortable giving good information while they are in pair, as compared to westerners, who have an individualistic approach (Yeo 1998)?

Therefore the evaluator and the test user relations need attention. What kind of evaluator and test user pairing will help us best to identify usability problems? Is cross cultural evaluator-test users pairing more productive? E.g. will evaluators from a western culture evaluating test users from an eastern culture be more productive? Or will evaluators sharing the culture of the test users be more fruitful in identifying cultural specific usability problems?

The answer to these questions and to the broad question about cultural effects on usability testing may be depend on the type of interface that is under consideration (e.g. localized interface vs. internationalized interface), the demographic characteristics of test users (e.g the young versus the old generation in India may dramatically differ in the computer skills) and the type of task that the test user is asked to carry out with the application to be tested (e.g. making an invitation with a word processor may be a very different task in a European culture versus a Chinese culture). We suggest, however, that these issues are treated as secondary to the primary problem of understanding how to ensure an efficient evaluator-test user relation in cross cultural Think Aloud usability testing.

The key issue in cross cultural Think Aloud usability testing, as we see it, is the relation between evaluator and test user and how it will vary the number of culturally specific usability problems. There are two kinds of relationships of possible significance:

1. The usability test relationship. One in which relationship is created by the evaluator (evaluator) with the test user during the course of the test so that the user shouldn't feel uncomfortable during the test and should provide a good quality of think-aloud-thoughts that allow the evaluator to identify important usability problems. This is the classical relationship discussed the theory and addressed by most guidelines for practical usability testing.

2. The contextual relationship. Another type of relationship, in which the relationship already was there between the evaluator and the test user because the evaluator and the test user were members of same culture, family, friends, good colleagues or in other ways already more or less strongly tied to each other. This is a relationship that needs to be discussed if we are to do cross cultural Think Aloud usability testing.

What we therefore want to investigate experimentally is

1) is the contextual relationship important for the usability test performance and outcome, given a 'standard' usability test relationship?
2) is the local evaluator – local test user relationship the most effective, i.e gives the best outcome in the fastest time, of the different contextual relationships?

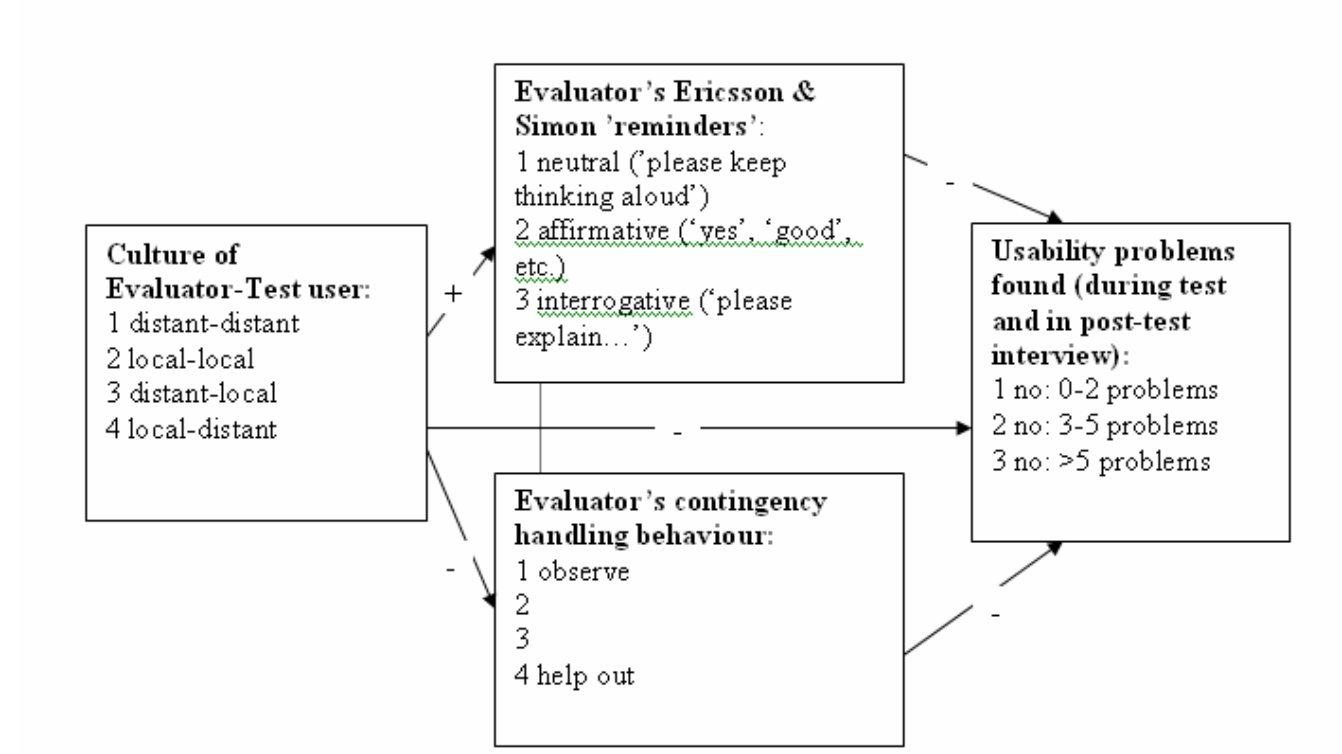

Figure 1. Explanatory model of cultural influences on outcome of Usability Test.

' + ' = positive influence, ' - ' = negative influence

METHOD

We conducted a three phase experiment to explore the effects of culture on the interaction between evaluator and test user in a usability test, when the culture of evaluator is different from the user and when it's same, see figure 1.

The phase one was Questionnaire phase which gave us the idea about the expertise of the user (as Microsoft Word user) and evaluator (as a Usability engineer and as a Microsoft Word user). It also gave us an insight how these persons have changed due to the effect of intermingling with other cultures.

The second phase was Usability testing of a Cultural Clipart application with the help of Microsoft word; this was followed by interviewing the test user by evaluator. During this phase the researchers (the authors of this paper) studied the interaction between the evaluator and test user.

This was followed by third phase: the interview phase. In this phase the researchers interviewed the evaluator and test user on the basis of their observations during phase two.

The participants

Test users and evaluators were chosen from Europe (Denmark) and India. As we can't take east and west as areas of study because culture is not at all uniform in whole east or in whole west (Nisbett 2003), we choose one country from east and compared the result with one country from west, i.e. we chose national culture as a way to operationalize the selection of participants with differences in cultural practices.

Name	Role	Age	Citizenship
Carsten	Evaluator	52 years	Danish
Mikkel	Test User	26 years	Danish
Lene	Evaluator	45 years	Danish
Sudhanshu	Test User	34 years	Indian
Shivam	Evaluator	20 years	Indian
Toke	Test User	30 years	Danish
Ram	Test User	27 years	Indian

Table 1. Description of participant roles, age and citizenship

Furthermore, despite the risk of selecting culturally too homogenous participants, we wanted participants that

- Could do the experiment in English. We asked them therefore about their ability in English and their experience with and knowledge of usability testing and thinking aloud. They should be able not only to talk English, but also to read English and to work with computers in English. They should be able to Think Aloud and even to do this in a foreign language like English.
- Could use the application and could perform the test-task. Therefore, the participants should be regular users and proficient in the 'host application' for the specific application to be tested, i.e. Microsoft word, and preferably have experience with the clipart features, and they should have experience with the test user task, i.e. making a birthday invitation by using a word processor and clipart.

The participants' details are presented in table 1. Notice that one evaluator, Shivam, was used two times as an evaluator due to difficulties finding an evaluator from India staying in Denmark.

To check our immediate selection of participants, they were screened by having to give us the above presented information about themselves by having to answer a questionnaire. This questionnaire also served to help us in the cultural pairing of Evaluators and Test users shown in Table 2.

Evaluator- test user relationship			
Cultural Pairing	Status	Age Relation	Sex
European European	Assistent. Proff – PhD. Student	Older-Younger	Female- Male
Indian Indian	PhD. Student – bachelor student	Young- Young	Male- Male
European Indian	Research ass. – PhD. Student	Older–younger	Male- Male
Indian European	Bachelor student – master student	Young-Young	Male- Male

Table 2. Participants' relationships.

As a substitute for established acculturation indexes (see eg (Zheng et al. 2004)), we tried to get a rough measure of test users' familiarity with local national culture. We asked the participants about their time of their stay in the local national culture (Denmark), which should be either from birth (local test user) or between two months and six months (distant test user; at least two months stay to avoid test users with cultural shock and rejection of local culture known to occur as a

response immediately after the ‘honeymoon experience of being in another country). Local national culture was thus expected to be either very familiar or to be very unfamiliar (but interesting) to the test users. We did not compensate for status or age and sex relations, but focused on conducting all the possible cultural pairings combinations of test user and evaluator pairs.

The test application

The interface that we have chosen was itself culturally localised. It was cultural clipart application, which is collection of culturally specific images and icons of various cultures. It also contained tools which can help user to edit those images. This application will help user to made invitations, documents containing graphics, web pages.

The need for localizing clipart collections. At present people use The Clip Organiser feature provided by the Microsoft Office quite often. But the problem that we found with the clip organiser is that it’s not currently localised, meaning the images and graphics that it contains now don’t have the cultural sensitivity for each culture in particular. Our intention is to make “Cultural Clipart” application which would be more sensitive to culture.

The prototype cultural clipart collection. Since our application was on the ideation phase we combined it with Microsoft Clipart. We have incorporated our Cultural Clipart in this clip Organiser by adding a sub folder named Cultural Clipart in My Collections. The Cultural Clipart folder contains the cultural specific sub-folders named Denmark, Germany, Sweden, etc. From any of these folders, user can choose images and graphics to add to their document. Our basic aim of these tests was to identify relationship between user and evaluator and how their familiarity affects the result of the test and not to identify Usability problems.

The creation of a cultural clipart collection. How did we get all the clip arts as the experimental material? The clipart material was a result of a two-week long camera-ethnography on the streets of Copenhagen + discussions about Danish culture with an anthropologist + search for clipart on the internet+ analysis of examples of Danish birthday invitations. In the end, we went for 'tourist' clipart to symbolize Denmark, realizing that the task we designed for the test user required such rude cultural distinctions.

In order to be able to evaluate whether the clipart used in this experiment can be matched and equivalent to typical clipart useful for performing a similar task in other cultures, e.g. in China or India, it is necessary to consider that our clipart collection had the following characteristics:

a) the number of clipart in total was 143

c) the number of task specific (birthday) clipart was, and

d) the number of deliberately introduced potential usability errors was 9. These errors are presented in the following.

Inclusion of potential usability problems. In order to increase the chances of interaction and that too cultural specific interaction between test user and evaluator we introduced some errors in the clipart whom only a user from Danish can only identify. We introduced the following usability problems in our Denmark folder.

- A Norwegian flag in the collection
- An image of Norwegian parliament in the Denmark folder
- Image of Heineken beer (a Dutch beer)
- A Reindeer (which does not live in Denmark)
- A Norwegian skier
- Giving the blind-fold game (not Danish) the birthday keyword
- Images of Birthday certificate, eagle and scenery of which none was Danish
- Wrong keywords to images of Danish flags
- Giving the keyword Birthday to an amusement park of Denmark
- Giving the keyword of birthday cake to various cakes that were neither birthday cakes nor Danish cakes

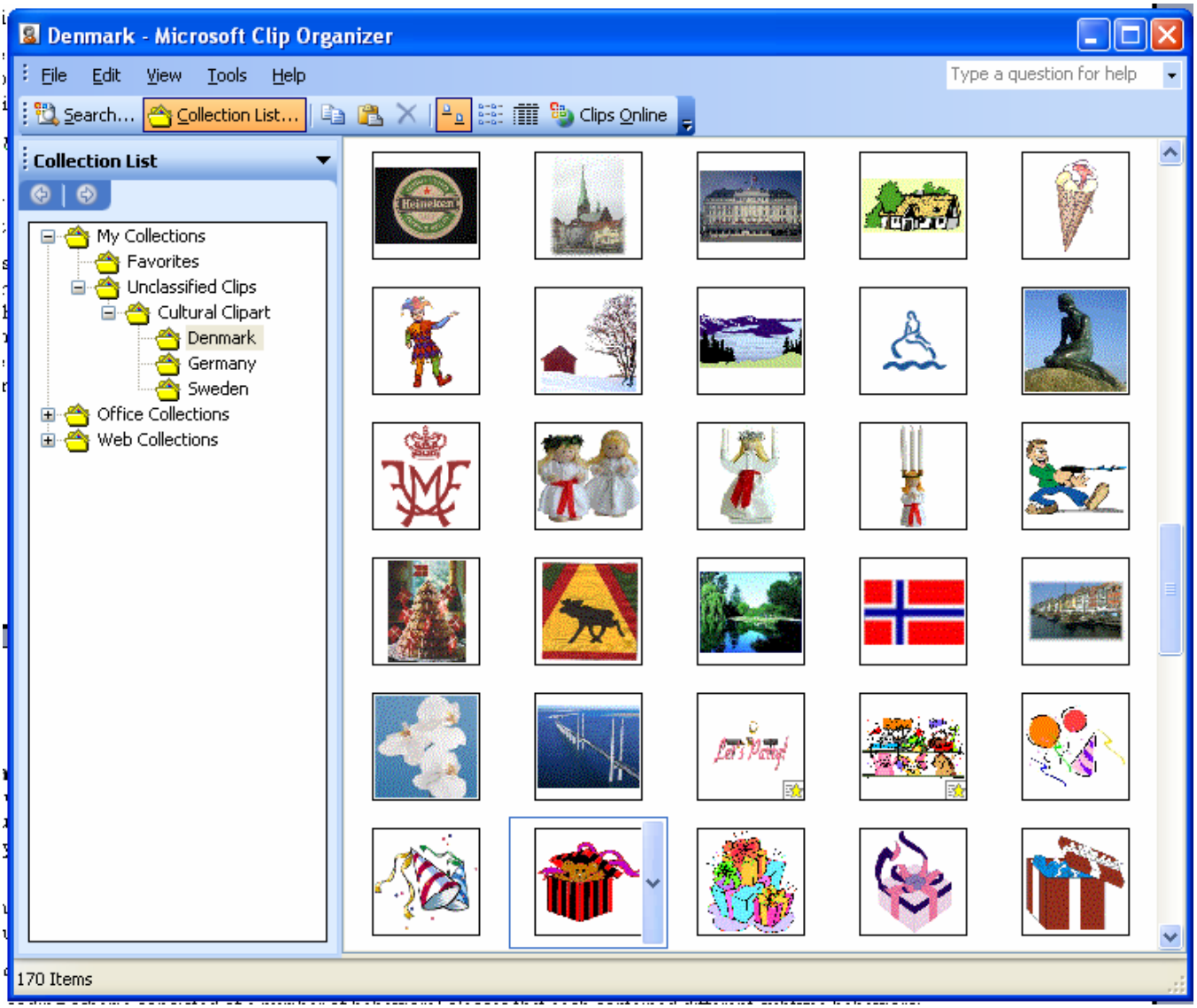

Figure 2. Denmark Cultural Clipart folder with built-in errors

The test scenario and tasks

In the test scenario, a company that develops a new application has hired a evaluator from a usability consultant company to perform Think Aloud usability tests of the product and has also provided test users. The test users and the evaluators were informed about the scenario and asked to provide background information about them in a questionnaire sent to them before the experiment took place. The experiment began then with the evaluator arriving at the company's office, i.e. the research lab, followed by arrival of test user, test session, evaluator's post test interviews with the test user, and, finally, the experimenters' interview with both the participants before they left the lab.

Evaluator's task. Evaluators were asked to conduct test in the way they are comfortable and they have to build their own test plan and interview guidelines. We provided them with list of tasks that they need to ask user to perform. They also have been provided with objectives of the test, which was – for the evaluator – to identify usability problems in a broad sense, both regarding content and navigation of our cultural clipart collection.

Each test session began with the evaluator greeting the test user. The evaluator then explained the test set up, about the objectives of the test, about the interface, about video recording and camera.

Then the evaluator asked the test user to read all the sub tasks very carefully and then perform them.

After the completion of all the tasks the evaluator interviewed the user about the interface and the task he/ she performed. The questions of the interview were solely on Evaluator's choice and according to his test plan keeping in mind the objectives of the test.

Test user's task. The test user was provided with the task of making a birthday invitation for his son. It had the following sub tasks:

1. Please write the text that you want to appear on the Invitation.
2. Please choose the appropriate font(s) for the text.
3. Please choose the appropriate style(s) for the text.
4. Please choose the colour(s) for the text.
5. You are free to choose any kind of formatting and layout that you require for this text.
6. Now using the Cultural Clipart sub-folder in My Collections folder in Microsoft Clip Organiser add some images and graphics so that its looks like Birthday invitation.
7. Please make this invitation look happy, colourful, and joyful as this is for birthday.
8. Since primarily all your guests are from Denmark and are Danes, make this invitation look Danish.

Test user was asked to read all the sub tasks very carefully and then perform them.

The test users performed the tasks using Microsoft Word (Microsoft Office XP Professional) on a Windows 2000 computer.

After the completion of all the tasks the evaluator had to interview the user about the interface and the task he/ she performed. The questions of the interview were solely on Evaluator's choice and according to his test plan that keeping in mind the objectives of the test.

Data collection

The experiment was performed as a laboratory experiment and took place in our research lab, which basically was a standard office space that could have been found in any major company.

Data collection

Both the test user and evaluator were asked to fill in a questionnaire which judged their knowledge of Usability testing and Microsoft Word. The questionnaire also gave an insight into their knowledge of the two cultures.

We recorded all instances of the participants' behaviour during the Think Aloud period from three different angles. We had a digital video camera directed at the test user's chair allowing capture of the test user's facial expression, and another digital video camera placed at 2 m distance allowing a capture of evaluator and test user interaction, as well as their interaction with the experimenter during the post test interview. Furthermore, we used the Camtasia ® screen recorder software to capture the screen events. Among these three sources of data, the primary source was the wide-angle camera that allowed us to see both participants as their relationship unfolded during the experiment.

The data collection method was thus continuous, i.e. test participants were observed for the whole period of the Think Aloud usability test, including evaluators introduction and post-test interview with the test user, but excluding the researchers' post-session interview with the evaluator and test user, see the test plan in appendix 1.

The research interviews.

The Think Aloud session was followed by three research interviews. The researchers interviewed first the evaluator and test user at the same time, and, secondly, the researchers interviewed the evaluator alone and, thirdly, the researchers interviewed the test user alone.

These interviews were conducted to explore the relationship between the evaluator and test user during the test, e.g. the interviewer probed into the interaction between the test user and evaluator; and how the user felt during the test and what else he/she wanted from the evaluator and did the nature of the evaluator's reminders affect his or her performance, and, furthermore, did the evaluator feel that he or she understood the test user's thoughts. The test user and the evaluator were asked to relate to the other's statements in order to get a dialogue about their relationship and its effect on the Think Aloud during the usability test and the number of usability problems that were found.

Data analysis

We wanted to produce a complete behavioural record of the participants' relationship as it unfolded during the experiment, including the time at which each instance of a behavioural relation occurred (events) or began and ended (states). Our analysis of the videos therefore focused on the wide-angle camera that allowed us to see both participants.

The continuous data collection allowed the both the frequency and the duration of different behavioural relations to be coded from the video. As equipment for coding the video, we used the Observer 5.0 software on a x86 Family 6 Model 8 Stepping 6 Genuine Intel with 254 MB memory and Windows XP Professional Edition Service Pack 1.

The coding scheme

The coding scheme consisted of a number of behavioral classes that each contained different subtype behaviors:

Reminders: Neutral (Ericsson & Simon type reminder), Affirmative (active listening type reminder), Interrogative.

- **Interrogative** Reminders are those in which Evaluator asked a question to make the Test user think aloud. E.g. What are you trying to do? What are you looking for? Etc.
- **Affirmative** Reminders are those in which Evaluator conveys a message to the Test user that he/she (the Evaluator) is an active listener. They could be easily looked as Boren and Ramey Acknowledgement tokens. E.g. Hmmm, Mm Hmm, Yeah, Ok.
- **Neutral** Reminders are those which are given to ask test user to continuing thinking aloud. E.g. Keep Talking, Please Keep Thinking Aloud, What are you thinking? (In a non-interrogative tone)

Task fulfillment evaluator behavior of different types: Observe Silently, Comment, Answer Questions, Help Out.

- **Observe Silently** means the Evaluator acted as a passive listener and didn't say anything while user was thinking aloud, or asking question or having difficulties.
- **Comment**s are those in which evaluator passed some comment E.g. The Evaluator has to interrupt and say that "You have to read the task first", "You can only use Microsoft Word to fulfil this Task"
- **Answer Questions** are those evaluator behaviours in which he/she has to answer, user's doubt about either the task or the application.
- **Help Out**s are those behaviours in which Evaluator actually helped user to complete the task. These are the strangest behaviours which are against any Usability Testing theory.

Usability problems related to: understanding image content, finding clip organizer, modifying images, choosing images and other problems, including general use of the word processor.

- **Images:** These are the number of responses which user gave when he was not happy with image collection, quality, and its cultural significance.
- **Clip Organiser:** These are the number of responses which the user gave when he had problems using the Clip Organiser. E.g. Name of the Image couldn't be found the Keywords were not right, it doesn't show the actual size of the image etc.
- **Choosing:** These are responses given when he/she had troubled transferring the file from the clip organiser to the Word Document.
- **Word:** These are the problems related to the Microsoft word.

User behavior of different types: Thinking Aloud, Silence, Explanation, Positive Comment, Negative Comment, Cultural comment, Question, Suggestions, Other responses.

- **Suggestions:** These are the responses given in order to improve the usability of the application.
- **Cultural Comments:** These are the responses and references made by the user on the users' native cultural or on the localised culture of the application.

- **Positive Comments:** These are the user responses which said about the positive-ness of the design and application.
- **Negative Comments:** These are the user responses which commented upon the disapproval of design or the application.

These codes were applied on the videos by one of the authors; however each class and subtype of behaviour was the decision of both authors.

Further analysis

Elementary statistics were performed on the codes, as presented in the result section below. The analysis of the interviews was used to interpret and backup our understanding of the results of the coding.

One major limitation with our approach was our focus on events, i.e. how many times a test user began to think aloud. Instead, it would have been relevant for comparison between subjects to get a measure of states, i.e. how long the evaluator and test user was in a 'Think Aloud mode', as the important thing is to get the user to Think Aloud much of the time and not just many times.

RESULTS

This section presents the result of the analysis of the video in tables and provide initial interpretations in brief form of the results.

Reminders (Given by Evaluator to User)

The European evaluator in the European-European pair was most affirmative to the responses given by the user as compared to any other pair. In this case the evaluator was particularly using the 'affirmative' acknowledgment tokens as suggested by Boren and Ramey as an alternative to the classical reminders.

Table 1. Reminders: The table shows the total numbers of the different kinds of reminders given by Evaluator to the test User in the four cultural pairings. **Note:** The test Session Duration represents the time taken by the user to complete the task. It doesn't include the follow up interview session time.

Evaluator-Test user pair	Test Session Duration	Interrog-ative	Affirm-Ative	Neutral	Total
European-European	26 min.	16	51	2	69
European-Indian	30 min.	13	7	10	30
Indian-European	26 min.	25	28	5	58
Indian-Indian	45 min.	18	13	9	40

The total numbers of reminders were least in the case of European-Indian test pair; also this evaluator was least affirmative in his responses to the user.

Task Fulfilment Evaluator's Behaviour

The result clearly showed that the European-European Test pair had interacted the most during the test session whereas there was very little interaction between Indian-European test pair.

Table 2. Task fulfilment Evaluator's Behaviour TFEB: The above table shows the evaluator behaviour in a test session. The number here represents the number of times each behaviour occurred. **Note:** Test Session Duration represents the time taken by the user to complete the task. It doesn't include the follow up interview session time.

Evaluator-User Test Pair Culture	Test Session Duration	Observe Silently	Comment	Answer Questions	Help Out
European-European	26 min.	44	3	7	9
European-Indian	30 min.	56	7	5	3
Indian-European	26 min.	60	1	5	0
Indian-Indian	45 min.	109	3	4	0

Notice that the European evaluator in European-European test pair after knowing that the user haven't used the Clipart before started helping the user. The other European evaluator in European-Indian pair didn't help his/her user that much.

It's also interesting to see that the comments made by the evaluator were least in the case of Indian-European pair; this could be argued as an unfamiliarity issue.

User Responses

User responses are those responses which are given by user during the course of the test. These responses could be either suggestion, problems cited with the interface, some comments made or in general thinking aloud. The number here represents the number of times each response was said. They are not the unique number of responses.

Level 1 and 2 usability problems

As shown in table 3, the European test user coupled with the European Evaluator responded most frequently on the usability issues concerning with images. Since the application is localised we interpret this result as meaning that an European test user will find more usability problem when coupled with European evaluator (they problems they found where knowingly introduced in the Clip Organiser by us, as explained above in the methodology section).

Table 3. Usability Issues (Level 1 and 2): These are the responses on the design of the interface. These are the first response they talk aloud if they had any kind of difficulties with the Images, Clip Organiser or related to Microsoft Word

Evaluator-User Test Pair Culture	Usability Issues (Level 1 and 2)			
	Images	Clip Organiser	Choosing	Word
European-European	5	2	0	5
European-Indian	1	1	0	5
Indian-European	0	4	0	4
Indian-Indian	1	3	1	15

Level 3 usability problems

As illustrated in table 4, one of the interesting thing that we found was that when an Indian user was asked about what a Norwegian Flag is doing in a Denmark folder (This was intentionally done) He said "Oh its perfectly fine…because in my opinion Norway and Denmark share a common culture so its perfectly fine…. ". Thus, even though the two test user that were European had the same status, age and experience, the European test user in the local distant relation commented less on Usability Issues called Images. European-Indian pair didn't make as many cultural comments as Indian-Indian pair, and, furthermore, the Indian Users asked many questions. Generally our result showed that the most foreign pair also had most cultural comments.

Table 4. Usability Issues and the Interaction between Evaluator and user (Level 3): These are user responses from their past knowledge about both the culture and the interface itself. But we have said this as level 3 data as they are users' immediate response about the interface but something coming from their long term memory. But we have included this into the list of usability issues as they will enhance the usability of the product.

Evaluator-User Test Pair Culture	**Usability Issues and the Interaction between Evaluator and user (Level 3)**				
	Test Session Duration	**Sugge-stions**	**Cultural Comments**	**Positive Comments**	**Negative Comments**
European-European	26 min.	5	4	3	2
European-Indian	30 min.	3	7	3	4
Indian-European	26 min.	0	6	3	2
Indian-Indian	45 min.	5	8	3	9

General TA behaviour

As shown in table 5, the European test user with the European evaluator was the most active think-aloud test user given the time spend on the session, although the Indian test user with the Indian evaluator had more think aloud events in total.

Table 5. General TA: These are the normal Think aloud responses meaning they were scored like user begin Think Aloud e.g. 40 number of times after the break in Think Aloud and it also contains the silence responses meaning the number of times the user went silent. This also contains the interrogative responses from the user which hindered the normal Think Aloud procedure.

Evaluator-User Test Pair Culture	**Thinking Aloud**			
	Test Session Duration	**Questions**	**Silence**	**TA**
European-European	26 min.	11	12	40
European-Indian	30 min.	16	5	27
Indian-European	26 min.	7	14	32
Indian-Indian	45 min.	12	18	48

DISCUSSION

In this section, we present the main findings from the pilot study and discuss additional findings and observations.

Firstly, having a local evaluator testing local users is significantly faster than control conditions. Why having a local evaluator testing local users is significantly faster than other combinations of evaluators and test users cannot be explained from standard text book procedures for TA (Ericsson et al. 1993) that basically says that all combinations are equally good, and neither from current approaches to cultural cognition (Nisbett et al. 2001) that talks about differences in cognition between people from US and China/Japan, but does not take into account the diverse background and languages of Europeans and Indians. Therefore we may have to look for an explanation in other perspectives on cultural differences; thus in a 'culturalization' perspective the local-local relation is most effective in finding usability problems in a localized application because of the unique characteristics of the target culture, while in a 'cultural representation' perspective the local-local relation is most effective in finding usability problems because of the meaning the evaluator and the user are able to ascribe to the localized application (Dormann et al. 2002). Which explanation is optimal under which condition remains an important topic for further research.

Secondly, a local test evaluator give significantly more reminders to a local test user than are given in control conditions, in particular more affirmative reminders, resulting in a significant higher amount of test user thinking aloud events. This result underlines that a TA usability test is far from being a standardized test in the sense of for example an international psychological intelligence test. In the study of such standard intelligence tests, researchers analyze the tests cross-culturally, based on the standardization data of thousand of individuals from several countries globally and investigate hypotheses on the degree to which there are universals in cognitive processes across cultures and variations in the cognitive processes due to specific cultural influences, and the differences in scores across cultures on specific subtests or scales of the test. The results showed a remarkable validity and reliability across cultures[i]. This remarkable result, however, were created by linking ecological and sociopolitical contexts to psychological variables, instead of determining cultural dimensions on basis of the aggregation of psychological data (Georgas et al. 2004). In the pilot study presented here, however, the approach was to explore the cultural representations in the concrete, psychological interaction between humans and computers. Therefore, the TA usability test in the present study was considered a cultural context in which to study the evaluators' and the test users' relations with each other and with the technological artifact, and the usability problems with the localized artifact were thus associated with the relations between concrete evaluators and test users in specific contexts of use.

Secondary findings

Finally, there are a number of secondary findings from the pilot study related to the possibility of outsourcing usability work and the general discussion of introspection in a cross cultural perspective. We discuss these in the following paragraphs.

Local evaluators help out. A local evaluator help out local users significantly more times when they are in trouble with their task or the equipment, he or she answer questions from the test user significantly more and fewer times observe silently when the test user is in trouble.

Local evaluators find task-specific usability problems. A local evaluator identifies significantly more task-specific usability problems with local test users than are found in control conditions. This is a new finding which has not been reported hitherto in the literature. (Molich et al. 2004) explicitly stated that in their comparative study of different usability teams: 'There didn't seem to be any relation between the number of test participants and the number or type of the reported problems', (Molich et al. 2004, p.72). Contrary to what Molich's statement says, the findings from the pilot study suggest a cultural-model explanation of the usability problems found that says it is possible to predict the type of usability problems that will be found (local-local will find more task-specific problems!)

The number and type of usability problems found appear also to be related to the complexity of the application under study (Molich et al. 2004, p.72). If the application to be tested is too large in terms of screen pictures or otherwise to be covered by one test, then the test will not find all usability problems no matter the kind of relation between test user and test evaluator. Local evaluators working with local test users may however cope with considerable more complexity and still find the interesting usability problems, because of their more intimate knowledge of the task to be used in the test.

Local evaluators give fewer cultural comments. There is a tendency that the local evaluator gives fewer cultural comments than is given in control conditions, or rather that in the control conditions many cultural comments are given. This tendency to comment on differences in cultural background seems to be a fixed ingredient wherever cross cultural tests are given.

Local evaluators inspire test users to think aloud. The test user Think Aloud with a local evaluator significantly more than in control conditions. As stated above, this may have to do with the type of reminders given by the evaluator (affirmative reminders). It remains however to be seen whether the high amount of thinking aloud is due to this particular type of reminder or whether the type of reminder that give many think aloud events vary with culture, i.e. the cultural model for 'encouraging thinking aloud' varies. For example, (Yeo 2001) noticed that in Malaysia English is mixed with the local

language - bahasa melayu - and that this makes statements more difficult to understand and to interpret, even for local evaluators. Similar conditions may be true in India – English may be mixed with Hindi or any of the local languages – and in Europe, for example when English is mixed with Danish in Denmark or with both Danish and the local language in Greenland.

Next step

These results underline that a TA usability tests are far from standardized tests in the sense of for example psychological tests. Central issues such as the implementation of the methodological approaches vary across evaluator-test user pairs.

Our next step is to conduct an empirical study with expert test users that perform a task in the presence of and expert test evaluators from more than two different cultures, i.e. from more than Europe and India. In particular, we also want to add evaluators with a background in cultures from China, as there is ample evidence that there are fundamental cultural differences in cognition between eastern and western cultures, e.g. while people from western countries think the world is a line, people from China think the world is a circle (Nisbett 2003). Experimenting with combinations of evaluators and users from different countries will, we hope, help us to propose cost effective and culturally sensitive Think Aloud usability evaluation techniques.

CONCLUSION

Experience has shown us that international usability testing of localized applications is best done by using local evaluators. This finding is confirmed by the results of the pilot study presented in this paper. The results allow us to suggest more detailed and explicit explanations of the factors behind this the finding. The results show that in a Think Aloud Usability Test think aloud usability test of a localized application is the most effectively performed when both the effective relation between evaluator (evaluator etc) and the test users is that both are local. Given the use of standard think aloud procedures in all experimental conditions, the local-local relation has major at least two big advantages compared to the control conditions. Firstly, having a local evaluator testing local users is significantly faster than control conditions. We have no text book procedures for TA that might explain why having a local evaluator testing local users is significantly faster than other combinations of evaluators and test users. However, the results may better interpreted in a ‘cultural representation’ perspective than a ‘culturalization’ perspective (Dormann et al. 2002). Secondly, a local test evaluator gives significantly more reminders, in particular more affirmative reminders, to a local test user than are given by the test evaluator in control conditions, which produce in particular more affirmative reminders, resulting in a significant higher amount of test user thinking aloud. The results underline that TA usability tests are far from standardized tests in the sense of for example psychological test user identifies, compared to control conditions, significantly more usability problems with the localized application. Central issues such as the implementation of the methodological approaches vary across evaluator-test user pairs.

Limitations of the this study

The limitations of this study are related to our choice of English as a common language for all participants. Obviously, not all users of technology speak fluently English. For example, for a hindi-speaking population there may be specific aspects of the Think Aloud Usability Test that are significant for the performance of the test and translation and transfer or these findings is a separate problem that are not dealt with in this study. Furthermore, villagers in rural areas of India may be not only unable to think aloud, but also believe that verbalization is a weird practice that is negative and deify both a local and a distant test evaluator as an authority figure. The ‘cultural model’-approach used in this pilot study may however also be applied to shed light on these barriers for the use of Think Aloud Usability Tests cross culturally.

ACKNOWLEDGMENTS

We are thankful to Human Computer Interaction Research Group at the Department of Informatics, Copenhagen Business School and Tom Plocher, Honeywell, for their kind support and help.

REFERENCES

1. Anderson, C. "How much interaction is too much?," *STC Usability Newsletter*) 2004.

2. Boren, M.T., and Ramey, J. "Thinking aloud: Reconciling theory and practice," *IEEE Transactions on Professional Communication* (43:3), Sep 2000, pp 261-278.

3. Bourges-Waldegg, P., and Scrivener, S. "Meaning the central issue in cross cultural HCI design.," *Interacting with Computers* (9) 1998.

4. Chavan, A. "Welcome to the Global Village: Some Considerations for Doing Usability in the Global Markets," *UI Design Update Newsletter* (March) 2004.

5. Dormann, C., and Chisalita, C. "Cultural values in web site design," The 11th European conference on cognitive ergonomics ECCEII Catania, Italy, 2002.

6. Ericsson, K.A., and Simon, H.A. *Protocol Analysis. Verbal reports as data* Cambridge Massachusetts, 1993.

7. Ericsson, K.A., and Simon, H.A. "How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking.," *Mind, Culture, & Activity* (5:3) 1998, pp 178-186.

8. Georgas, J., Vijver, F.J.R., and Berry, J.W. "The ecocultural framework, ecosocial indices, and psychological variables in cross-cultural research," *Journal of Cross-Cultural Psychology* (35:1), January 2004, pp 74-96.

9. Herman, L. "Towards Effective Usability Evaluation in Asia: Cross-cultural differences," OZCHI 1996, 1996.

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

11. Hertzum, M., and Pejtersen, A.M. "The information-seeking practices of engineers: searching for documents as well as for people," *Information Processing & Management* (36:5), Sep 2000, pp 761-778.

12. Kessner, M., Wood, J., Dillon, R., and West, R. "On the Reliability of Usability Testing," CHI2001, 2001.

13. Khaslavsky, J. "Integrating Culture into interface Design," CHI 1998, 1998.

14. Law, E., and Hvanneberg, E. "Analysis of Combinatorial User Effects in International Usability Tests," CHI 2004, 2004.

15. Marcus, A., and Gould, E. "Cultural Dimensions and Global User-Interface Design: What? So What? Now What?," 6th Conference on Human Factors and the Web, 2000.

16. Miller, M.D., and O'Donnel, C. "International Usability Testing: How Can we Do It Early, Often and Cost-effectively?," CHI 1994, 1994.

17. Molich, R., Ede, m.R., Kaasgaard, k., and Karyukin, B. "Comparative Usability Evaluation," *Behavior and Information Technology* (23:1), January- February 2004, pp 65-74.

18. Nielsen, J. "Designing for International Use," CHI 1990, 1990.

19. Nielsen, J., Clemmensen, T., and Yssing, C. "Getting access to what goes on in people's heads? - Reflections on the think-aloud technique," NordiCHI 2002, Århus, Denmark, 2002, pp. 101-111.

20. Nisbett, R.E. *The Geography of Thought* Nicholas Brealey Publishing, London, 2003.

21. Nisbett, R.E., Peng, K.P., Choi, I., and Norenzayan, A. "Culture and systems of thought: Holistic versus analytic cognition," *Psychological Review* (108:2), Apr 2001, pp 291-310.

22. Russo, P., and Boor, S. "How Fluent is Your Interface? Designing for International Users," INTERCHI 1993, 1993.

23. Schneiderman, B., and Plaisant, C. *Designing the User Interface*, (Fourth edition ed.) Pearson Addison Wesley, 2004.

24. Shweder, R.A., and Bourne, E.J. "Does the Concept of the Person Vary Cross-Culturally?," in: *Thinking Through Culture - Expeditions in cultural psychology*, R.A. Shweder (ed.), Harvard University Press, London, 1991.

25. Smith, A., and Yetim, F. "Global human-computer systems: Cultural determinants of usability. Editorial.," *Interacting with Computers* (16) 2004.

26. Vatrapu, R. "Culture and International Usability Testing: The effects of Culture in Structured Interviews. Master thesis.," in: *Virginia Polytechnic Institute and State University*, 2001.

27. Vatrapu, R., and Pérez-Quiñones, M. "Usability Testing: The Effects of Culture in Structured Interviews," 2004.

28. Vöhringer-Kuhnt, T. "The influence of culture on Usability. Master thesis.," in: *Dept. of Educational Sciences and PsychologyFreie Universität Berlin, http://userpage.fu-berlin.de/~kuhnt/thesis/results.pdf, July 2004.*, Berlin, Germany., 2002.

29. Yeo, A. "Cultural Effects in Usability Assessment," CHI 98, Doctoral Consortium, 1998, pp. 74-76.

30. Yeo, A. "Global-software Development Lifecycle: An Exploratory Study," CHI 2001, 2001.

31. Zheng, X., Sang, D., and Wang, L. "Acculturation and subjective well-being of chinese students in Australia," *Journal of Happiness Studies* (5) 2004, p 57–72.

32.

Test Plan

Total Test Participant Time - 1 Hour 20 Minutes

Total Evaluator Time - 1 Hour 40 Minutes

Total Time to conduct one Test - 2 Hours 10 Minutes

	Time	Activity	Details
Notes	15 min	Set up/ Prepare the Lab	Fix all the equipments, Cables, Sitting Arrangements, Cameras, Softwares, Etc.
	10 min	Welcome Evaluator	Test Organizer will welcome the guest Evaluator and give him/ her the Following Details: 1. Brief Introduction About the Thinking Aloud Technique 2. Test Goals 3. Task Specifications and Details 4. Task Length 5. Interview Guidelines 6. Instructions to Conduct the Test
	10 min	Material Study	Evaluator will study all the material given to him and specially tasks.
Video	15 min	Welcome Test Participants	Evaluator will brief him/ her about the Thinking aloud, the interface going to be tested, the purpose of test and what is expected to come out of the test, etc.
Continuous recording of behaviour to be coded later	15 min	Perform task and Think Aloud	Test participant will now Think Aloud as he/ she performs the task. He/ she will be provided with different sub task as he/ she completes the previous sub task.
Video	15 min	Interview/ Debriefing Session	Evaluator will now interview the Test Participants
Video	30 min	Interview Session 2.1(Both) Interview Session 2.2(User) Interview Session 2.3(Evaluator)	In Interview Session 2.1 Test Organizer will interview both evaluator and test participants. In Interview Session 2.2 and 2.3 Test organizer will interview the evaluator and test participants individually.
	5 min	Good-bye and Thanks giving	
	15 min	Cleaning Up the Lab	

[i] As an illustration of how far the standardization of intelligence tests has proceeded, three abstracts of papers published at the 28th International Congress of Psychology, Beiijing, 2004, are presented below:

Cross-cultural psychology, intelligence and cognitive processes, **J. Georgas**, The University of Athens, Athens, Greece

The WISC-III was analyzed cross-culturally, based on the standardization data of 15,999 children from 16 countries: USA, Canada, Britain, Austria, Germany and Switzerland, France and French Speaking Belgium, the Netherlands and Flemish Speaking Belgium, Greece, Sweden, Lithuania, Japan, South Korea, and Taiwan. The hypotheses were: (1) the degree to which there are universals in cognitive processes across cultures, (2) the degree to which there are variations in the cognitive processes due to specific cultural influences, and (3) the degree to which there are differences in mean scores across cultures on the WISC-III subtests, FSIQ, VIQ, PIQ, and Index scores.

The WISC-III: History and contemporary cross-cultural perspectives, Georgas, J. , Weiss, L.G. , van de Vijver, F.J.R., & Saklofske, D.H. (Eds) (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III*. San Diego: Academic Press., University of Saskatchewan, Saskatoon, Saskatchewan, Canada

This presentation will outline both the history of the development of the Wechsler test for assessing children's intelligence and the current use of these tests in 16 countries. The influences and experiences that led David Wechsler to produce the Wechsler Bellevue scale in 1939 will be examined. Factors that resulted in subsequent scale revisions culminating in the WISC-III will be described along with an analysis of current status of this test. An overview of the standardization studies and clinical use of the WISC-III in the 16 countries included in our recent research study will allow for a cross-cultural evaluation of the test. Of particular relevance are the changes required to adapt and standardize the WISC-III in culturally and linguistically unique countries while preserving the integrity of the test.

A cross-cultural analysis of the WISC-III, **F. van de Vijver**, Tilburg University, Tilburg, The Netherlands

This paper presents a cross-cultural analysis of the WISC-III, based on 12 standardization data-sets from 16 countries (N = 15,999). The WISC-III showed a remarkable similarity in factor structure across the countries. The next analysis compared the Subtest scores, the factor-based Index scores, and the Full-Scale IQ, Verbal IQ and Performance IQ across the countries. A comparison of the mean scores showed remarkably small differences. The final analysis explored the relationships between these average scores at the country level and the ecocultural indices of Affluence and Education. Education showed slightly smaller correlations with IQ scores than Affluence did.