Evaluation Methods and Cultural Differences: Studies Across Three Continents

Cecilia Oyugi
Thames Valley University
St Mary's Road
Ealing W5 5RF, London, UK
+44 (0)208 280 0258
cecilia.oyugi@tvu.ac.uk

Lynne Dunckley
Thames Valley University
St Mary's Road
Ealing W5 5RF, London, UK
+44(0)208 280 0258
lynne.dunckley@tvu.ac.uk

Andy Smith
Thames Valley University
St Mary's Road
Ealing W5 5RF, London, UK
+44 (0)208 280 0256
andy.smith@tvu.ac.uk

ABSTRACT

This paper reviews issues and problems that arise in cross-cultural usability evaluations. It reports two separate empirical studies of a number of well-known techniques with UK, African and Indian users. The studies examine the effectiveness of methods based on think-aloud protocols, including the DUCE method, to elicit users' views. The results from all the studies show that these established Western methods are less effective with users from other cultures. It suggests that the reasons for this are the consequences of deep-rooted differences in personal interactions in different cultures. This paper provides evidence to guide choices for applications involving users from India and Africa.

Categories and Subject Descriptors

H.5.2 User Interfaces, J.4 Social and behavioral sciences

General Terms

Measurement, Performance, Design, Human Factors,

Keywords

Cross-cultural evaluation. usability methods, international usability evaluation.

1. INTRODUCTION

Although ICT researchers and practitioners have long been aware of the challenges of the global market, there are still many unsolved problems concerning the extent to which culture may affect the development of the artifacts produced [18]. Research into cross-cultural user interface design has established the existence of a cultural effect in the development and use of ICT which goes beyond language differences. For example, studies based on usability metrics clearly show differences in performance with users from diverse cultures [5] although the reasons for these differences can be difficult to pin down without insights into the users' thinking. At present the origins and consequences of these cultural effects remain controversial.

NordiCHI 2008: Using Bridges, 18-22 October, Lund, Sweden
 ISBN 978-1-59593-704-9. $5.00

However there are two key bodies of research: one which emphasizes and extends usability principles to other cultures [2] and another which emphasizes the different context of evaluators and users [21, 12, 1]. In relation to the *product* of development, cultural differences in signs, meanings, actions, conventions, norms or values raise challenging issues in the design of usable localized artifacts. In relation to the *process* of development, cultural differences potentially affect the manner in which users are able to participate in design and act as subjects in evaluation studies. Local people will have their own concepts of knowledge and their own forms of information communication so that it is essential that they should be able to shape their use of ICT without the risk of losing their culture and identity.

As long ago as 1996, Herman [11] noted that the results of user-based testing indicated that cultural effects exist and exert a strong influence on the outcome of user interface evaluation. In addition he recognized the need to modify 'Western' usability evaluation methods for application in the Far East. Since then a number of other researchers have reached the same conclusions [6, 23, 24] particularly in relation to evaluation methods that seek to elicit users' attitudes through the use of think-aloud methods and structured interviews.

Although these evaluation methods were grounded in a user-centred approach, the users and developers have distinct roles and separate contributions that they can make to the design process. It is the user who experiences the system, interacts directly with the design factors that determine usability and benefits from the usability characteristics of the system. Users however are not experts in HCI and are not able to analyse or articulate directly their requirements for the interface. Murphy [14] documents some of the problems that can arise in international usability testing.

Several different sources of these problems have been put forward ranging from simple translation errors to deeper cultural misunderstandings. These have been interpreted as the effects of cultural dimensions such as power distance and uncertainty avoidance that lead to misunderstandings between users and evaluators or indeed between computer professionals themselves. As a result cross-cultural user evaluations can lead to invalid conclusions or data which is difficult to interpret.

Users vary in their ability and willingness to articulate their thoughts to the evaluator depending on both their individual personality and cultural background. One approach has been to modify 'Western' methods by using different task scenarios. For example, recognizing that in many Asian countries the main challenge with usability testing is that it is impolite to tell someone they have a poor design Chavan [3] has developed the Bollywood method (which inherits from the Bollywood film genre, which typically involves 'emotionally involved plots with great dramatic flourish) to reduce user inhibitions.

In order to help clarify the issue the authors have been undertaking on-going research into evaluation methods. This paper describes investigations of a number of related user evaluation methods in two 'non-Western' cultures and contrasts the behaviour and results obtained from UK users with those from India and Africa. In the following section the user evaluation methods are outlined in terms of their different rationales before descriptions of the two investigations are given.

Capturing the user's immediate experience is attempted by a number of related methods that are variously known by the terms think-aloud, verbal protocol and co-operative evaluation. All are linked by the concept that the best way to understand users' experiences of an interface is to observe people as they operate and concurrently talk about their experience. However, there is evidence that the actual process of *thinking aloud* can cause the subject of the study to proceed differently and can lead users to encounter different usability problems. Henderson et al. [10] have investigated the latter issue, focusing on variations of the technique, including:

- *thinking aloud*, in which the users 'think aloud' while doing the task,
- *record and thinking aloud*, in which user actions are recorded, later replayed and users are then asked to explain their actions.

Cooperative evaluation is a variation of *thinking aloud* in which the user is encouraged to see himself as a collaborator in the evaluation rather than just a subject. This is claimed to be less constrained and the user is encouraged by an evaluator, who is not necessarily the designer, to actively criticize the system. These 'traditional' methods of testing are difficult to operate across cultures and remote geographical locations [4].

Another variant of these methods called DUCE (Designer User Contextual Enquiry) has the objective of making the user explain their normal working practice in relation to the prototype and while they are interacting with it [16]. Users are required to interact with high-fidelity prototypes using task scenarios drawn from their working practices and are asked to verbalise their experience. In order to assist the user to do this the evaluator is required to ask the user a number of open questions as interaction progresses. The philosophy behind the user questioning within DUCE is not unique. For example questioning is also a feature of Co-operative Evaluation [22], however the questioning style is more exploratory and less inquisitorial, for example questions in the style of 'why did you do that' are excluded because it was felt this would make the evaluator too dominant in the conversation. The advantages of DUCE are that it combines methods that would be situated and that would be feasible for early interactive prototypes and redesign of existing interfaces. Secondly DUCE facilitates the identification of specific design improvements from the usability data collected.

Cultural differences in usability evaluations are multi-dimensional. As discussed above, they can occur as a result of cultural differences inherent *within different cultural user groups*, with groups potentially reacting differently to individual evaluation methods. Differences can also be evidenced as a result of cultural differences *between users and evaluators*. In relation to the latter issue Vatrapu and Pérez-Quiñones [21] present a controlled study investigating the effects of cultural differences between Indian users and evaluators from different cultures. The results showed that participants found more usability problems and made more suggestions to an evaluator who was a member of the same (Indian) culture than to the foreign (Anglo-American) interviewer.

2. EXPERIMENTAL STUDIES

The two dimensions of cultural differences *within different cultural user groups*, and *between users and evaluators* may interact, making an even more complex environment for analysis. Therefore, designing a methodology for studies in comparative cross-cultural usability involves four different user-evaluator study options:

(a) users from both cultures resident in one culture, with one single evaluator [5]

(b) users from both cultures resident in different (home) cultures, with one single evaluator [6]

(c) users from both cultures resident in one culture, with two different evaluators, with evaluators and users from the same culture [21]

(d) users from both cultures resident in different (home) cultures, with two different evaluators [14]

All options have problems. Option (a) is the most simple to undertake but acculturalization of the non-native user group could reduce any cultural differences between the two groups. Acculturalization is defined as the social and psychological integration of individuals with the target language group [19] and occurs as the dominant host culture absorbs to a certain extent minority immigrant culture [20]. Option (b) addresses the acculturalization problem, but not the problem of user-evaluator cultural differences. Options (c) and (d) have the problem that different evaluators may approach the methods in different ways, thereby introducing a further variable affecting the results. In these studies options (a) and (b) were followed.

2.1 Indian Case Study

In order to further examine cultural differences with Indian users a study was undertaken to compare three different evaluation methods with Indian users: 'post-usage interview', 'think-aloud only' and 'think-aloud with probing'.

Having reviewed methodology options and also considering logistical issues it was decided in this study to select Option (a) as described above. Users from both cultures (India and UK) were resident in one culture (UK), and one single evaluator (from India) was used in all studies. In order to address the acculturalization problem Indian users were selected after measuring their individual acculturalization levels using the Suinn-Lew Asian Self Identity Acculturalization (SL-ASIA) scale

(Suinn, Ahuna and Khoo, 1992). Only individuals with low acculturalization levels were selected. All evaluation work was undertake in English both in India and the UK. Indian users' natural language for business activities such as this was English.

Five users from each culture were selected for each of the three methods ('post-usage interview', 'think-aloud only', and 'think aloud with probing'). There were therefore two sets of 15 users from each culture. Two websites were used for each study, one from UK and one from India. Usability evaluation sessions were video recorded and analysed after each session. Within the analysis all user feedback and categorized as either *useful information* (something relating to the suability of the system) or *non-useful information* (something un-related to usability, such as content based information). The number of items of *useful information* about the usability of the web site derived from each user was recorded.

2.1.1 Results of Indian Case Study

The detailed results show differences in performance of the UK and Indian users for each method as summarized in Table 1. A cursory analysis would indicate that Indian users generated fewer items of useful usability related information than UK users with all three methods (67 compared to 82, 130 compared to 144 and 171 compared to 258).

Table 1. Total and mean number of items of information captured from each user

	Post-usage Interview		Think aloud		Think aloud + probe	
	UK	Indian	UK	Indian	UK	Indian
Total	82	67	144	130	258	171
Mean	16.4	13	29.2	26	51.6	34.2

In this case the evaluator was Indian and although Vatrapu and Pérez-Quiñones [21] suggest that users normally provide more information to an evaluator who was a member of the same culture the Indian users provided fewer items with all three methods. These results would initially suggest that these methods are indeed less suitable for Indian users.

Table 2 summarizes the results for the three methods used in terms of the number of individual items of usability information captured from each evaluation for all thirty users, with descriptive statistical data. It is clear that 'think-aloud with probing' is the most effective method overall and this elicits more than three times as much useful information as the 'post-usage interview' method.

Table 2. Three Methods Compared

Method	N	Mean	Standard Deviation	Min	Max
Post-usage Interview	10	14.7	6.54	6	26
Think aloud	10	27.6	7.13	15	38
Think aloud + probing	10	42.9	11.92	30	66
All methods	30	28.4	14.51	6	66

The results were then analysed in more detail to check the significance of these differences. The type of experimental design determines the analysis of variance procedure that should be adopted. In this case an independent measures design was adopted where the dependent variable scores are assumed to be statistically independent or uncorrelated, i.e. the subjects were allocated randomly to each method and each subject provided one dependent variable score (the number of items of useful usability information). In order to statistically measure the effect of the Culture Factors (India, UK) and Method Factors ('post usage', 'think-aloud', 'think aloud + probing') the data was analyzed using the General Linear Model (GLM) for independent measures using SPSS. The results summarized in Table 3.

Table 3. Summary of Analysis of Variance for two factors, Culture and Methods

Source	Sum of Squares	Df	Mean Square	F	Sig.	*Eta Squared*
Culture	480.00	1	480.00	8.77	.007	*.268*
Method	3985.80	2	1992.90	36.40	.000	*.752*
Culture * Method	331.40	2	165.70	3.02	.067	*.201*
Error	1314.00	24	54.75			
Corrected Total	6111.20	29				

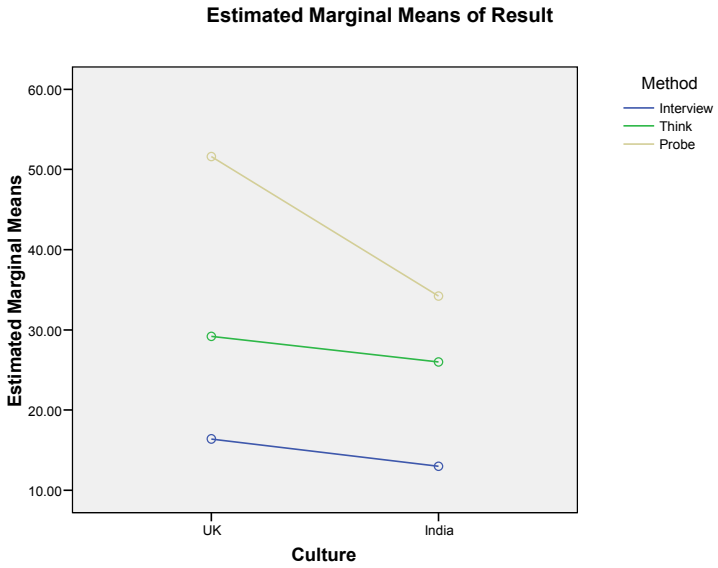

Figure 1. Differences in means for different methods and cultures.

Although ANOVA is a powerful method that has been used extensively for analysis of variance, it is claimed that ANOVA assumes that the underlining populations are normally approximately Gaussian (especially with large samples). On the other hand non-parametric methods such as Krusal-Wallis are less suitable for significance test involving small samples. It was for this reason that it was decided to use the General Linear Model (GLM) to analyse the data since it is based on creating a model which is presented as the GLM equation which does not require the data to be normally distributed, merely that the errors are normally distributed.

Table 3 results shows that the model based on the two factors is statistically significant and the Culture Factor and the Method Factor make a significant effect on subjects' performance. The null hypothesis relating to these factors can be rejected. However the interaction of the two factors Culture*Method is not statistically significant at the (0.05) level. This can also be seen from the graph shown in Figure 1. However it can be seen that the performance of the UK and Indian users was quite close for the first two methods. When checked individually it should be noted that these differences were not significant when analysed separately by One-Way-ANOVA (p 0.44 for 'post-interview', p 0.511 for 'think-aloud only'). The highly significant difference was for the 'thinking-aloud with probing' method. Note that since there are three methods to consider, the alternative approach would make three t-tests necessary. The problem with a t-test analysis is that the probability of a type 1 error (i.e. rejecting the null hypothesis when it is true) increases with the number of significance tests carried out. When the usual significance level of 0.05 is used when two-t-tests are applied it rises to nearly double the ANOVA tables significance level. In contrast ANOVA simultaneously examines for differences between any number of conditions while holding the type 1 error at the chosen significance level.

The partial eta squared statistics in Table 3 reports the "practical" significance of each term, based upon the ratio of the variation (sum of squares) accounted for by the term, to the sum of the variation accounted for by the term and the variation left to error. Larger values of partial eta squared indicate a greater amount of variation accounted for by the model term, to a maximum of 1. Here although the individual terms are statistically significant, the greatest effect is from the Method Factor.

2.2 African Case Study

The context for the African study is the VeSeL (Village e-Science for Life) project, which is part of the Bridging the Global Digital Divide Network funded by the UK's Engineering and Physical Sciences Research Council. The VeSeL project's objectives, which are focused on Kenyan rural farming communities, address both educational and technological issues and, subsequently, designing and testing appropriate technologies to meet their needs. The VeSeL project involves two community groups in rural Kenya while the developers are drawn from five UK universities. As part of the project one of the community groups requested a website to promote their projects such as the eradication of Tsetse fly with the hopes of attracting more funding from globally distributed users. An early prototype blog site was developed by researchers from the London Knowledge Laboratory. The blog contains basic information about the community group, their mission, vision and simple means of communicating with the group through comments and email. The blog site is intended for global users who might be interested in donating funds to the community.

In view of the results from the Indian study described earlier, *that highlighted differences between different cultures involving probing,* it was decided to explore the DUCE method which had been used successfully with many UK commercial developments [16] but had not been used for cross-cultural evaluation.

One of the features of the DUCE method is the questioning style. It is suggested that each DUCE session should be treated flexibly, however questions are provided and structured under Norman's Seven Stages of Action [15]. These can be asked at appropriate points within each discrete user action. The overall process is presented in Figure 2 and potential questions are provided in Table 4.

For each task / goal
Ask the user to explain what he / she is attempting
For each sub task
Ask the user to explain what he / she is attempting
For each stage in Norman's model of interaction
Consider asking a question from the check list
Next stage
Next sub task
Next task

Figure 2. Eliciting user comments in a DUCE session

Table 4. Check list of questions

Norman's Stage	Potential Question
1 Form a goal	a) How does the screen help you select a way of achieving your task?
2 Form an intention	b) How does the screen suggest that what you are about to do is simple or difficult
3 Specify the action sequence 4 Execute the action	c) How does the system let you know how you are making progress?
5 Perceive the resultant system state 6 Interpret the resultant state	d) What is the most important part of the information visible now? e) How has the screen changed in order to show what you have achieved?
7 Evaluate the outcome	f) How do you know that what you have done was correct? g) How would you recognise any mistakes?

One of the issues of concern was that the DUCE questions were developed and tested in a Western cultural context and it was uncertain whether the same style of questions would be effective with African users. Since researchers [21] have reported that the richness of data obtained from evaluations using structured interviews was influenced by the cultural match of evaluators and users the DUCE method was investigated with users from the UK and Kenya based on Option (b). The users from two cultures – UK and Kenya - participated in their own cultural environment with one evaluator who was Kenyan but had been acculturalized to UK culture. It was expected that DUCE would provide a rich source of user evidence that could be brought to bear on the enhancement of prototype user interfaces.

Although the Culture Factor made less contribution than the Method Factor it was still statistically significant particularly in the case of the commercially most popular method, 'think-aloud with probing'.

2.2.1 Method for African Case Study

The evaluation was carried out on two sets of users, Kenyan based users and UK based users, who carried out seven tasks using the blog site. The users, five from each culture, were tested in their home cultural environment and were matched by age, gender and levels of education background. Both sets of users were familiar with the Internet and were drawn from socio-technical groups who were likely to donate online. This was similar to the socio-economic background of the Indian and UK users in the first study. They were not paid and were not informed about the experimental design and hypothesis. The evaluator had a Kenyan background and was the same for both sets of users. She ran the evaluation sessions as a regular usability evaluation and each session was audio recorded. The evaluator was the same for both sets of users. The evaluator's laptop was used during all sessions. All evaluation work was undertaken in English both in Kenya and the UK. Kenyan users' natural language for business activities such as this was English.

The overall hypothesis was that there would be no difference in the type of response obtained from both sets of users using the DUCE method. The user was then given one task at a time and asked to carry it out. During task execution, the evaluator asked questions using the DUCE format.

2.2.2 Results of African Case Study

Elicitation of information from UK users was relatively easy and the feedback obtained was more detailed as can be seen from Table 5 which shows a typical user's response for one task - *You want to tell the community members that the Tse Tse fly project is commendable. Post a comment on their blog website.*

Table 5. UK user's response for Task 4

What UK user said	What user did
" I cant find anywhere I can post a comment. Do I have to give my user name and password before I can post my comment?"	Chose anonymous option
"Wheelchair access-am not sure what it means "It is not clear what the icon means. It has to do with wheelchair access doesn't make sense in computing. May be a sign for help. I wonder if there are other clearer icons that would be used. What does it mean in this context.	Clicked on the wheelchair symbol
"The system doesn't give feedback about my progress" Of course my post is posted nothing specific? "Yes, the comment has been posted, the feedback was at the very top of the page and I need to scroll up the page although I did not see it immediately."	Typed the comment and posted it in the web blog

As can be seen in Table 6, in the case of the Kenya-based users, elicitation of information was more challenging.

Table 6. Kenyan user's response for Task 4

What Kenyan user said	What user did
"I don't know how to post a comment on the web blog"	
"There is a post a comment link here, I will click on it and if it gives me somewhere to type then I will type my comment"	Clicked on the post a comment link and typed his comment alright.
"I don't think I have been able to publish my comments"	Since the user did not identify himself, when he clicked on the publish comment, this was not successful.

In order to check the results statistically analysis of variance was carried out using the GLM method as shown in Table 7. However although this shows that culture was statistically significant the Levene test was significant suggesting that the samples did not have equal variance, therefore the Kruskal Wallis Test was used and this confirmed that the null hypothesis that culture made no difference to the method should be rejected.

Table 7 Analysis of Variance

Source	Sum of Squares	df	Mean Square	F	Sig.
Culture	82083.60	1	82083.60	13.55	.006
Error	48468.00	8	6058.50		
Total	241222.00	10			

Far more items of information were elicited from the UK users than from the Kenyan users. In addition the Kenyan users were not comfortable with the probing questioning style of the DUCE method.

Several of the Kenyan users expressed uneasiness/irritation with the summary questions. The users commented that the evaluator was asking the same question in 7 different ways, and were fed up by the end of the exercise. In addition the evaluator felt:

(i)The users perceived the evaluation exercise as a test on them and for every task that was incomplete; they perceived it as personal failure.

(ii) Previous experience with ICT had a significant effect. Whilst both UK and Kenyan user groups were similar in composition with a similar spread in ICT expertise, those in Kenya with less experience found that the tasks were not as straightforward as did similarly experienced users in the UK.

(iii) Part of the DUCE method is a set of questions that seem similar but are nevertheless different. However, the users responded to these set of questions with constant irritation as it seemed as if one question was being asked 7 different times. This section of the evaluation was eventually not carried out because the users resulted in being frustrated and angry before the exercise was over.

Task complexity appeared significant. The first three tasks were relatively straight forward with the last being more challenging. The UK based users responded to the difficult task phlegmatically with many suggestions for improvements. In contrast when the Kenya users experienced failure with Task 4 the amount of information elicited dropped significantly and their frustration with continuing to answer the DUCE questions was noticeable.

3. DISCUSSION

All evaluation methods analysed here aim to gain meaningful information about the user's interaction experience. The results of both studies support the view that different cultures respond differently to evaluation methods and that 'Western' methods are much less effective in other cultures. In addition these studies do not support the conclusion [21] that the problem can be simply solved by using an evaluator from the same culture as the users. The UK users provided more information than the users of non-Western cultures even with Indian and Kenyan evaluators, whilst both the Kenyan and Indian users performed poorly with 'Western' methods with evaluators from their own culture.

With the Indian users, 'Think Aloud with Probing' is better than the other methods (Table 2) and is shown statistically significant (Table 3). Of course this Indian study does not give us any reason why such a result may occur. In the late 1990's HCI researchers and practitioners turned to cultural models to improve design, particularly Hall [9] and Hofstede [12]. Hofstede's work may be relevant here because of its focus on human interaction where cultural differences clearly matter. He carried out a study of 116,000 IBM employees distributed through 72 countries using 20 languages in 1968 and 1972. The study was based on a rigorous research design and systematic data collection [12]. He conceptualised culture as 'programming of the mind', meaning that certain reactions were more likely in certain cultures than in other ones, as a result of differences between basic values of the members of different cultures. Hofstede proposed that cultures could be defined through four dimensions :

- power distance -the degree of dependence between boss and subordinate (PD)
- collectivism versus individualism integration into cohesive groups versus being expected to look after him/her self (IND)
- femininity-masculinity -the extent to which gender roles are distinct or overlap (MAS)
- uncertainty avoidance -the extent to which members feel threatened by uncertain or unknown situations (UAI)

The use of Hofstede's dimensions to frame interface design remains controversial. However in the absence of more appropriate models we can reflect on the differences between the Hofstede dimensions for the UK, India and East Africa and make judgments as to whether his model is consistent with our findings as shown in Table 8. (Hofstede's analysis for East Africa includes the countries of Ethiopia, Kenya, Tanzania, and Zambia).

The largest differences in Table 8 between the UK and the other cultures appear to relate to Power Distance and Uncertainty Avoidance. To understand the sources of these different responses to these methods we can review the known problems with the methods. One problem is that a significant factor in the maintenance of human-human dialogue appears to be the expertise of the participants which could relate to Hofstede's PD values.

Table 8. Values of Hofstede's Cultural Dimensions

Group	PDI	IND	MAS	UAI
UK	30	83	61	35
India	77	50	56	65
East Africa	64	27	35	52

Falzon [7] describes dialogues between experts and non-experts (e.g. patient-doctor) which are analogous to developer-user situations where the expert speaker soon assumes control of the conversation and the remaining exchange of information follows a sequence of 'yes/no' questions and answers. While the UK users are unaffected by perceived power distances between their role as a user and the evaluator this difference may be the significant underlying cause of problems with this style of conversational method with the other two cultures. Another related problem is highlighted by Goguen [8] who criticizes 'think-aloud' methods as 'unnatural' for the reason that language is intrinsically social; it is created for a conversational partner. As a result a person imagines a partner with certain desires and tries to address these desires, at the expense of accuracy and reliability. In Table 8 the collectivism versus individualism scores (IND) show considerable difference between the UK with a high individualism score than with East Africa or India. This again could account for a reluctance to criticize which results in fewer usability issues being identified.

Lin et al [13] in their study of usability methods make the point that thinking aloud seems very unnatural to most people and some test users have great difficulty in keeping up a steady steam of utterances as they use the system. Inexperienced users find difficulty in verbalising their operations so that both the users and the evaluators need training for the technique to be effective.

We can also speculate about the effect of the differences in Uncertainty Avoidance. We have found evidence that Indian users have some difficulty in adapting readily to highly structured task-based testing. Both the interview and think-aloud only methods, being ones without evaluator interruption, allowed much more flexibility in user's interpretation of the required tasks. With the probing method their interaction is far more interrupted and this may inhibit flexible interaction. Continuing the speculation, this is potentially in accordance with India's supposed polychronic culture as defined in cultural models [9], in which multiple tasks are handled at the same time, and time is subordinate to interpersonal relations. It is possible that 'think-aloud with probing methods' reinforce monochronic interaction.

Hall [9] also contrasted high-context cultures such as India and Kenya and low-context cultures such as the UK and USA. These differences can be expressed by differences in locus of control, with high-context cultures (e.g. eastern cultures and those with low racial diversity) tending to inner locus of control with attribution for failure and personal acceptance for failure while low-context cultures tend to outer locus of control and blame of others for failure. This could explain the responses of the Kenyan users who saw their failure to complete the tasks as a personal issue rather than a failure of the usability of the blog site. Whereas for the UK users the opposite was true and the blame was placed on the design. Another feature of Hall's analysis is the use of non-verbal communication so that much more nonverbal communication is expected in a high-context culture while a low-context culture would have more focus on verbal communication than body language. This would mean that evaluation procedures should take into account the developers/evaluators body language, gesture, facial expressions and behaviour.

Initial indications from this experiment show that the DUCE method which was prepared within the Western cultural context may not necessarily be suited to another culture, in this case the African culture. There may need to be adjustments to the DUCE method so that the feedback to be obtained is as expected. In terms of the VESEL project this is important as it intends to identify and develop the most appropriate technologies including: radically different user interfaces for illiterate or semi-literate user groups.

4. CONCLUSIONS: FUTURE WORK

Culture remains difficult to study, alone and certainly in relation to HCI practices. It is particularly difficult to identify meanings, attitudes and expectations, not to mention the deeply embedded values and beliefs behind people's thoughts, behaviours and actions. Behaviours might be influenced by other factors (e.g., environmental conditions) rather than cultural traits, and the reasons for, and meaning of, an action can seldom be observed wholly and directly. Even so, studies of this sort are continuing to add to the knowledge bank of cross-cultural usability. In both studies reported here we have demonstrated further evidence that a greater understanding of cultural differences in both the process and the product of website development is necessary to ensure systems success.

The results reported here require us to consider the degree in which the replication of Western approaches to usability methods in India and Africa is to be encouraged. The problem is particularly acute as evidence in collaborative institutional projects in India and China [18] is that the developing local usability communities are probably too keen to implement 'best practice'; from the West before fully testing its relevance in the local culture. HCI practitioners need to develop other evaluation methods that are more appropriate for different user groups and complex and mobile interfaces. We must be prepared to explore different methods even though this may be challenging.

REFERENCES

[1] Abdelnour-Nocera, J., Dunckley, L. and Sharpe, H. (2007), An Approach to the Evaluation of Usefulness as a Social Construct International Journal of Human Computer Interaction, 2007, Vol. 22, No. 1-2, 153-172.

[2] Del Galdo, E.M (1996) Culture and Design. In E.del Galdo and J.Nielsen (Eds.) International User Interfaces. John Wiley and Sons, Inc. 74-87.

[3] Chavan, A. (2004), The Bollywood Method. in E. Schaffer,. *Institutionalization of Usability; a Step-by-Step Guide.* New York: Adisson Wesley. 129-130.

[4] Dray, S. (2001) Usable for the World: A practical guide to International User Studies in D. Day and L. Dunckley (Eds.) Designing for Global Markets 3, Proceedings of IWIPS 2001, pp 154-155. IWIPS.

[5] Dunckley L, Smith, A & Howard, D (1999). 'Designing for Shared Interfaces with diverse user groups.' INTERACT 99, pp 630.636. Eds. M.A. Sasse, C. Johnson. Chapman & Hall. 1999.

[6] Evers, V. (2001) Cultural Aspects of User Interface Understanding: An Empirical Evaluation of an E-learning Website by International User Groups. PhD Thesis, Open University.

[7] Falzon, P (1990). Human-computer interaction: Lessons from human-human communication. In P. Falzon (ed.) Cognitive Ergonomics (pp 51-68). London: Academic Press.

[8] Goguen, J. A. (1996), Formality and informality in requirements engineering, Proceedings of the second international conference on requirements engineering, (ICRE'96), IEEE Computer Society Press, pp 102-108.

[9] Hall, E. T. (1976) *Beyond Culture*, Doubleday, Garden City, New York.

[10] Henderson, R.D., Smith, M.C., Podd, J., Varel, A. and Alvarez, H. (1995) A Comparison Of The 4 Prominent User-Based Methods for Evaluating the Usability of Computer Software. *Ergonomics*, **38**, 10, 2030-2044.

[11] Herman, L. (1996)"Towards Effective Usability Evaluation in Asia: Cross-Cultural Differences," *ozchi*, p. 0135, *6th Australian Conference on Computer-Human Interaction (OZCHI '96*).

[12] Hofstede, G.(2001) *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations.* Thousand Oaks CA: Sage Publications.

[13] Lin, H. X., Choong, Y. and Salvendy, G. (1997), A proposed index of usability: a method for comparing the relative usability of different software systems, *Behaviour and IT,* 16 (4 and 5), 267-278.

[14] Murphy, J. (2001) Modelling, designer-tester-subject relationships in international usability testing, in D. Day and L. Dunckley (Eds.) *Designing for Global Markets 3* pp 33-44. IWIPS.

[15] Norman, D. (1988) The Psychology of Everyday Things. Basic Books, New York.

[16] Smith, A. & Dunckley, L. (2002) Prototype evaluation and redesign: structuring the design space through contextual techniques. *Interacting with Computers*. 14 , DEC (6): 821-843.

[17] 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): 63-91.

[18] Smith, A., Gulliksen, J., and Bannon, L. (2005) Building usability in India: reflections from the Indo European Systems Usability Partnership, in T. McEwan, J. Gulliksen and D. Benyon (Eds.) People and Computers XIX - The Bigger Picture: *Proceedings of HCI 2005* Springer.

[19] Spolsky, B. (1989). *Conditions for Second Language Learning.* Oxford: Oxford University Press.

[20] Suinn, R. M., Ahuna, C. and Khoo, G. (1992) The Suinn-Lew Asian Self-Identity Acculturation Scale: Concurrent and Factorial Validation. *Educational & Psychological Measurement* 52(4), pp.1041-1046.

[21] Vatrapu, R. and Pérez-Quiñones, M. A. (2006), Culture and Usability Evaluation: The Effects of Culture in Structured Interviews, *Journal of Usability Studies*, Issue 4, Volume 1, pp. 156-170.

[22] Wright, P.C. and Monk, A.F. (1991) A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies*. 35 891-912.

[23] Yeo, A. (2000), Usability Evaluation in Malaysia. Proceedings of 4th Asia Pacific Computer Human Interaction Conference: APCHI 2000. Elsevier, pp 275-280.

[24] Yeo, A. (2001) Global Software Development Lifecycle: An Exploratory Study. In Jacko, J., Sears, A. Beaudouin-Lafon, M. and Jacob, R. (Eds.) CHI 2001: Conference on Human Factors in Computing Systems. ACM Press, pp 104-111.