Language Issues in Cross Cultural Usability Testing: A Pilot Study in China

Xianghong Sun[1] and Qingxin Shi[1, 2]

[1] State Key Laboratory of Brain and Cognitive Science, Inst. of Psychology, Chinese Academy of Science, Beijing 100101, China
sunxh@psych.ac.cn
[2] Department of Informatics, Copenhagen Business School, Denmark
qs.inf@cbs.dk

Abstract. Language effect (Chinese vs. English), and power distance between evaluator and user in usability test were investigated. 12 participants from China, Swede, and Denmark formed 7 evaluator-test user pairs. Test users were asked to use a software. Evaluators were asked to conduct the usability test, and try to find usability problems. Participants' conversation, behaviour, and screen operation were recorded by behaviour observation system. Results showed that Speaking Chinese made evaluator giving more help in detail, and encouraging users more frequently; Speaking English asked evaluator and user look at each other more often to make themselves understood, and evaluators paid more attention to check task list. Power distance also had effect on evaluators and users. When evaluator's title were higher than users, evaluator would pay more attention to users' doing, not like to give user detailed instruction, usually loose communication with user, and spent less for task management. In contrast, talking to evaluators with higher rank, users tend to use more gesture to express themselves.

Keywords: Language, think aloud, cultural usability, field study.

1 Introduction

With the progress of economic globalisation, more and more international enterprises start to do usability test in different cultures during the last decade. In China only two, or three years ago, usability was quite a new word for most of people. Right now the situation has been changed dramatically. Many domestic enterprises have reckoned the importance of usability test for their products, especially for IT business. Many western researchers were interested in Chinese users' preference, behaviour, and mental models [3,4,5,8]. Since China is not an English speaking country, like India, and Singapore, and most users in China can't speak English at all. It brings the biggest communication problems when conducting usability test by international moderators.

There are several choices to avoid this problem. The first is using bilingual moderators to test user. The second is finding users who can speak English. But both professional moderators and English speakers are very rare in China, and they all are

N. Aykin (Ed.): Usability and Internationalization, Part II, HCII 2007, LNCS 4560, pp. 274–284, 2007.
© Springer-Verlag Berlin Heidelberg 2007

youth and probably with western education background, it means that there is no way to get the real feedback from all kinds of users in China. So the third and the most regular way they do is that, they use both remote and local moderators working together with Chinese users to ensure they really get the feedback from the right users and understand it.

Local moderators here mean someone who got training in Human Factors, or had working experience on usability test for at least one year in China. They usually can't speak English very well. Remote moderators mean someone who got training in Human Factors and had experience on usability test for at least one year in foreign countries. They usually can speak English and their hometown language very well.

Previous studies on cross cultural usability evaluation show us that culture broadly affects the usability evaluation processes [9]. Vatrapu R, and Pérez-Quiñones M.A (2006)[10] investigated the evaluator effect, and found that participants found more usability problems and made more suggestions to an interviewer who was a member of the same (Indian) culture than to the foreign (Anglo -American) interviewer. The results of the study empirically establish that culture significantly affects the efficacy of structured interviews during international user testing.

In this study, the primary questions are how to 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? Before we can answer them completely, First thing we need to do is to find what kinds of cultural factors could affect usability test. In this paper we investigated specifically two factors: one was language, and the other was power distance.

The reason why we picked language as a factor to be investigated is that, language is a kind of representation of culture. And language situation among India, European countries, and China is totally different. Although English is not hometown language for Indian and Danish either, most people in these two countries can speak English very well. But in China few people can do it well. Therefore, if conducting usability test in China, First thing you have to do is to change the testing interface into Chinese. We usually say if someone is speaking English, he/she must be thinking in English. So, by which language test user and evaluator choose during the usability test, they probably think in the way of that language. It means that speaking different language could affect the process of usability test even if all the participants are Chinese.

Since China is a kind of society with very clear and strict hierarchy. Different kind of relationship between evaluator and test user could cause different results. So, power distance is considered as another factor.

2 Method

The following section will describe the methodology, results, and conclusion of the pilot study and discusses the findings on language issue.

2.1 Materials

The pilot study was based on the 'usability test of cultural clipart' paradigm (Clemmensen, 2005)[1]. Here cultural clipart was a collection of culturally specific images and icons and several text documents with preformatted invitation text. The application was aimed at supporting a test user in the design of invitations. In this study, test users were asked to make a wedding invitation for themselves.

Totally 150 images and icons with wedding symbols were selected and saved in a subfolder with the name “Chinese clipart” in My Collections, in which 20 image and icons with Korean and Japanese symbols and another 10 with western style were mixed with others as interference (see table 1) to increase the chances of measuring culturally specific interaction between test user and evaluator.

Test user could access the images in “Chinese clipart” folder with Microsoft’s clipart organizer.

Table 1. Potential usability errors in Chinese Cultural Cliparts

Culturally wrong symbols	*Label errors*	*Invitation text errors*
A Korea flag in the collection		Wrong time
An image of cherry flower (Japanese national flower)		Wrong place for wedding banquet
Image of bride in traditional Korean wedding dress		Wrong name and title
Japanese rope node		Wrong telephone number

2.2 Procedure

The pilot study in China had the same three phases like the other two experiments did in Denmark and India: phase one was Questionnaire phase which gave us the information about the experience of the user and evaluator; phase two was Usability testing of the Cultural Clipart application with Microsoft word. This phase included two parts, first was the testing, and second was interviewing the test user by evaluator. The third phase was the interview phase: the researchers interviewed the evaluator and test user on the basis of their observations during phase two.

The whole experiment was conducted at a standard usability lab in Institute of Psychology, which included one test room with several video camcorders installed in different viewpoints, and one observation room with one-way mirror between the two rooms. All the conversation between evaluator and test user, their behaviour, and the screen events were recorded by four-channel behaviour recording system.

2.3 Participants

Table 2 showed the basic information of all the seven evaluator-user pairs. Here all the test users were chosen from China. They are young staff, or graduate students studying in the Institute of Psychology, which ensure them all familiar with think aloud technology, and speak good English. The evaluators were chosen from Europe and China. Only one evaluator who was from Swede had not any knowledge about usability. But he got half an hour of training before he conducted the usability test. Since there is few people in China now could be treated as professional usability test leader, one of the evaluator was used three times with different test users.

Table 2. Description of participants

ID	*Role*	*Age*	*Gender*	*National culture*	*Language used in the test*	*Expertise in usability test*
E1	Evaluator	34	F	China	2 in English 1 in Chinese	Professional
U1	Test User	31	F	China	English	
E2	Evaluator	37	F	China	English	Professional
U2	Test User	25	F	China	Chinese	
E3	Evaluator	45	M	Danish	English	Professional
U3	Test User	27	M	China	English	
U4	Test User	29	F	China	English	
U5	Test User	27	M	China	English	
E4	Evaluator	25	F	China	Chinese	Non- professional
U6	Test User	24	F	China	Chinese	
E5	Evaluator	27	M	Sweden	Chinese	Non- professional
U7	Test User	23	M	China	Chinese	

Table 3. Different combinations of evaluator-user relationship

Cultural Pairing	*Status*	*Age Relation*	*Language*	*Gender*
Chinese Chinese	Prof. – Prof.	Young-Young	English – English	F- F
Chinese Chinese	Prof. – PhD student	Young-Young	Chinese – Chinese	F- F
Chinese Chinese	Prof. – PhD student	Young- Young	English – English	F- M
European Chinese	Prof. – Prof.	Older–younger	English – English	M - F
Chinese Chinese	Prof. – PhD student	Young - Young	English – English	F - M
Chinese Chinese	PhD student – PhD stud	Young- Young	Chinese – Chinese	F - F
European Chinese	Bachelor – PhD stu.	Young- Young	Chinese – Chinese	M- M

3 Results

Coding system. Watching evaluator and test-users' conversation and behaviour, it's found that, even for a single event, for example, silence, the duration of the event last were varied from several seconds to several minutes. But in the original coding system used in India and Denmark experiments, no matter how long the event lasted it was count as once. It brought us two problems. First, how long the event last could be treated as one event? Secondly, when we found the number of silence for one evaluator-user pair was higher than that of another pair, did it mean the former pair had more silence than the latter? It's probably not true. If the former pair fallen into silence 10 times during the test, and it last around 30 seconds each time, the total time in silence is 300 seconds. If the latter pair in silence twice, and 5 minutes each, the total time was 10 minutes. So the real situation could be opposite.

In order to eliminate the system error we encode, a chronological coding system was developed, by which we encoded the behaviour data by time period instead of by event. Each time period last 10 seconds. For example, if the usability test lasted 20 minutes, there would be 120 time points were coded. Therefore, times that an event happened during the test must be equivalent to how long that user or evaluator spent on that event.

Since the focus of our study was the process of usability test, especially the interaction between user and evaluator, in the chronological coding system, evaluator's conversation, evaluator's behaviour, test user's conversation, test user's behaviour, and screen operation were all coded. (See Appendix. Coding system)

Language Effect. The first row in Table 4 showed how many time points for each evaluator-user pair were coded. The third row illustrated which pair spoke Chinese (C), and which spoke English (E). Here the focused issue was whether for Chinese people speaking English, not native language – Chinese, would make the test process different. Since there was one European participant in the pair No.2, and pair No.3, we only do analysis with the other 5 Chinese-Chinese pairs.

Table 4. Classification of participant pairs and the test duration for each test pair

E-U pairs No.	1	2	3	4	5	6	7
Number of time points for coding	345	131	181	134	189	295	199
Test session duration (mins)	88	20	30	20	31	49	33
Language they used (Chin, Eng)	C	C	E	E	E	C	E
Status (at same level, or not)	Same	Same	Same	No	Same	No	No

Table 5 showed different content of evaluators conversation. From the data, we found most of time evaluators were keep silent. But in speaking Chinese condition, silence kept longer than speaking English. Numbers of reminder kept the same between the two conditions. Numbers of affirmative express, and answering user's questions didn't show any clear trend. But under Chinese condition, evaluators tend to give more help, tell more introductions in detail, and encourage users more frequently.

Table 5. Classification of evaluators' conversation

Language	Chinese		English		
	1	6	4	5	7
1 Affirmative express	12	8	52	1	16
2 Remind user keep thinking aloud	3	1	1	0	1
3 Tell user what is next step	**25**	**12**	2	7	17
4 Interrogative express	5	15	5	2	8
5 Answer user's question	33	4	3	1	18
6 Help out	**15**	**15**	1	9	0
7 Encourage user	**6**	**9**	2	0	0
8 silence	243	225	68	169	138

Table 6 showed different kinds of behaviour of evaluators during the test. It's found watching PC screen was the most behaviour. The second most behaviour was checking task list to find what was the next step, or to ensure everything was done. Chinese evaluators seldom expressed their thought with gesture. They didn't turn their face only to look each other with user either. But when watching PC screen they did have a look of the user. Comparing the two language conditions, evaluator when speaking Chinese didn't do so much on task management as they did when speaking English. And English condition asked evaluator and user look at each other more times to make them understood. In addition, under English condition, evaluators paid more attention to check task list.

Table 6. Classification of evaluators' behaviour

Language	Chinese		English		
	1	6	4	5	7
1 Turn face to user	0	0	0	**1**	**2**
2 Express himself with gesture	0	0	0	2	0
3 Watch PC screen	295	111	125	208	162
4 Task management	14	20	**61**	**33**	**15**
5 1+3	28	3	**3**	**34**	**6**
6 2+3	1	0	**0**	**9**	**10**
7 1+2	2	0	0	0	1
8 1+2+3	5	0	0	8	2

Combining the above two results we found speaking different language affected evaluators' behaviour. Speaking Chinese made evaluators easier to give help and more detailed instruction. And speaking English made evaluator and user have to look at each other more frequently to ensure there was no misunderstanding between them.

Table 7. Classification of test-users' conversation

Language	Chinese		English		
	1	6	4	5	7
1 evaluation	21	21	16	36	24
2 suggestion	0	0	0	3	0
3 explanation	0	1	0	1	0
4 question	20	3	7	9	25
5 description	115	72	87	116	57
6 confirmation	29	2	5	24	16
7 silence	159	15	73	99	73

Table 7 and Table 8 showed user's conversation and behaviour. From Table 8 it's found that, user didn't use gesture either, and seldom look back to evaluator. Different from Table 6, user never spent time on task management. That meant they didn't check what task should do next. They gave the responsibility totally to evaluator.

Conversation of user was classified into 7 types. From Table 7, we found that users didn't have so much silence as evaluators had, which was what users were supposed to do. Only one user made suggestion to the clipart organizer. Chinese users didn't explain how did he/she think, and why he/she picked this picture, not that one. What he/she spoke out mostly were what he/she was doing. So they just described their screen operations to evaluator.

Table 8. Classification of test-users' behavior

Language	Chinese		English		
	1	6	4	5	7
1 Turn face to evaluator	0	0	1	2	0
2 Express him/herself with gesture	0	0	0	0	0
3 Watch PC screen	338	86	188	271	156
4 Task management	0	0	0	0	0
5 1+3	7	1	0	15	0
6 2+3	0	38	0	2	33
7 1+2	0	3	0	2	2
8 1+2+3	0	6	0	1	6

Comparing the two language conditions, there seemed no difference exist on the amount of evaluations to Chinese clipart, amount of questions, descriptions, and confirmation. So for users, whatever language they spoke, it didn't affect their conversation content and behaviour.

Power distance. Reviewing all the participant pairs, we found two of them were student-student pair, another two of them were professor-professor pair, and the other three were professor-student pairs (See the last row in Table 4 and Table 3). The first four pairs were treated as at the same status level. In each pair there was not power distance exist. The last three pairs were treated as at different level of status. Professor as evaluator in think aloud session was at least one layer higher than student.

Table 9 and Table 10 showed the evaluators' behaviour at different groups. From the data in the two tables, we found if evaluator's title were higher than users, the evaluator would more like to ask user what was he/she thinking at that time, would not like to give user more detailed instruction, and didn't remind user so much on keep thinking aloud as evaluator at the same level with user did. In addition, from the Table 10, evaluator with higher rank would loose more communication with user, and spent less time for task management.

So, evaluator's status had affected the interaction between user and himself.

Table 9. Effect of relationship between evaluator and user in evaluator's conversation

Status	Same level				Different level		
	1	2	3	5	4	6	7
1 Affirmative express	12	5	49	1	52	8	16
2 Remind user keep thinking aloud	**3**	**7**	**1**	0	1	1	1
3 Tell user what is next step	**25**	**10**	**22**	**7**	2	12	17
4 Interrogative express	5	9	2	2	**5**	**15**	**8**
5 Answer user's question	33	1	10	1	3	4	18
6 Help out	15	2	3	9	1	15	0
7 Encourage user	6	0	3	0	2	9	0
8 silence	243	93	89	169	68	225	138

Table 10. Effect of relationship between evaluator and user in evaluator's behavior

Status	Same level				Different level		
	1	2	3	5	4	6	7
1 Turn face to user	0	6	4	0	0	1	2
2 Express himself with gesture	0	0	0	0	0	2	0
3 Watch PC screen	295	81	69	125	111	208	162
4 Task management	**14**	**11**	**65**	**61**	20	33	15
5 1+3	**28**	**17**	**23**	**3**	3	34	6
6 2+3	1	2	13	0	0	9	10
7 1+2	**2**	**7**	**2**	0	0	0	1
8 1+2+3	5	3	3	0	0	8	2

Table 11 and 12 showed the difference of users' behaviour between participants at same level group and at different level group. It illustrated that users in No.2 and No.3 gave more explanation to their evaluators, paid more attention to task management, and had more face-to-face communication with evaluators. But it didn't mean that user with the same title would give more communication and explanation than user with lower title would do. Since the evaluators in participants pair No 2 and 3 were from European countries, probably it's because they were foreigners, users in this two pairs had to explain more, and had more face-to-face communication.

But in the Table 12, there was a power distance effect showed in row 6: watching PC +communicate with gesture. When evaluator's rank was higher, users tend to use more gesture during the test session. What could be the reason? When we went back to review the videotapes, we found evaluators with higher rank would sit a little farther with user than they did in another situation, and users felt a little nervous when they talk to evaluator with higher rank. It could be the reason.

Combining the data from user and evaluator, we can say power distance affected not only e valuators' behaviour, but also the users. But in this study, there was another factor involved with the power distance factor.

Table 11. Effect of relationship between evaluator and user in user's conversation

Status	Same level				Different level		
	1	2	3	5	4	6	7
1 evaluation	21	7	28	16	21	36	24
2 suggestion	0	0	1	0	0	3	0
3 explanation	**0**	**3**	**5**	**0**	1	1	0
4 question	20	8	15	7	3	9	25
5 description	115	55	107	87	72	116	57
6 confirmation	29	8	16	5	2	24	16
7 silence	159	39	9	73	15	99	73

Table 12. Effect of relationship between evaluator and user on user's behaviour

Status	Same level				Different level		
	1	2	3	5	4	6	7
1 Turn face to evaluator	0	5	0	1	0	2	0
2 Express himself with gesture	0	1	0	0	0	0	0
3 Watch PC screen	338	98	143	188	86	271	156
4 Task management	0	**13**	**14**	0	0	0	0
5 1+3	**7**	**12**	**10**	0	1	15	0
6 2+3	0	1	10	0	**38**	**2**	**33**
7 1+2	0	0	1	0	3	2	2
8 1+2+3	0	0	3	0	6	1	6

4 Discussion and Conclusion

Comparing with the results in pilot studies in India and Denmark, we didn't find much data related to comments, especially to culture. Only when evaluator was from another country, not from China, users would explain more about his/her choices of pictures and icons (see Table 11). It implied that when evaluator and user came from same culture background, although the test session could go through very quick and smoothly, they could probably miss some cultural usability problems. Another reason could be that, users were not get used to the way of think aloud, especially in process of design. Five of seven users mentioned in the follow up session that they were not satisfied their wedding invitation design because of the time pressure. For this point, we need rethink the experiment design including the wedding invitation task, and the think aloud method.

Although there were three phases for each participant pair, we didn't analyse the amount of usability issues that evaluator found. That because in this study, one evaluator was used three times, the number of usability problems she found was cumulated test by test. So there was no way to compare them between different conditions.

Another reason for that, when we go back to the original videotapes, we found not any user found the invitation text error. Many user did noticed some culturally wrong symbols in images and icons, but they didn't mentioned them until evaluator asked them pick them up. So, after the think aloud session, when researcher asked users the reason why they couldn't find the culturally wrong things, they told us they were so concentrated to fulfil the whole task so that they missed out all the details. And sometime although they noticed the wrong picture, but what they were asked to do was find an appropriate one as a decoration, so they thought it's not necessary to point it out.

From the data shown in Table 7 and 8, there was not much behaviour difference whatever users spoke Chinese, or English. It probably didn't mean that there was no influence of language on users. As mentioned before, it might because of the task requirement asking users doing what they did not do usually. The effect of task difficulty might have impact on the effect of language.

So, let's look back to the users' think aloud behaviour (Table 7), and calculate the percentage of each kind of conversation content. We found Chinese user spent about 30% of time on silence, another 30% of time on description, the other 30% on other contents, such as asking question, and evaluate the interface. Ramey doubted the think-aloud method in usability test [6]. In this study, there seemed having the same question: do we think keep talking about 60% of test time mean a real think aloud? Do we think half of talking time were spent on describing what he/she is doing is really a think aloud? Is it true that what a person's doing is what the person's thinking? Do we need to give more training before we start the usability test session? All the questions need to be answered in further study.

Briefly we can make conclusions here that, speaking different language and power distance affected the process of usability test. Speaking Chinese made evaluator giving more help, telling more about introductions, and like encouraging users more frequently; Speaking English made evaluator and user look at each other more often to make themselves understood, and evaluators paid more attention to check task list. When evaluator's title were higher than users, the evaluator would pay more attention to users' doing, would not like to give user more detailed instruction, usually loose more communication with user, and spent less time for task management. In contrast, talking to evaluators with higher rank, users tend to use more gesture to express themselves.

Acknowledgments. This study was co-funded by the Danish Council for Independent Research (DCIR) through its support of the Cultural Usability project. Thanks to the people who gave us access and helped us in this study.

References

1. Clemmensen, T.: Community Knowledge in an Emerging Online Professional Community: The Case of Sigchi.dk. Knowledge and Process Management 11(2), 1–10 (2005)

2. Hornbæk, K.: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64, 79–102 (2006)
3. Marcus, A.: International and intercultural user interfaces. In: Stephanidis, C. (ed.) User Interfaces for All: Concepts, Methods, and Tools, pp. 47–63. Lawrence Erlbaum, Mahwah (2001)
4. Marcus, A.: Fast Forward: User-Interface Design and China: A Great Leap Forward. ACM, New York (2003)
5. Marcus, A.: Cross-cultural, global, and mobile user-interface design. In: HCI International: 11th International Conference on Human–Computer Interaction, Las Vegas, USA (2005)
6. Ramey, J., Boren, T. et al.: Does Think Aloud Work? How Do We Know? In: CHI, proceedings (2006)
7. Rau, P.-L.P., Choong, Y.-Y., Salvendy, G.: A cross cultural study on knowledge representation and structure in human computer interfaces. International Journal of Industrial Ergonomics 34(2), 117 (2004)
8. Shen, S., Woolley, M., Prior, S.: Towards culture-centred design. Interacting with Computers 16, 1–33 (2006)
9. Smith, A., Yetim, F.: Global human – computer systems: cultural determinants of usability. Interacting with Computers 16, 1–5 (2004)
10. Vatrapu, R., Pérez-Quiñones, M.A.: Culture and Usability Evaluation: The Effects of Culture in Structured Interviews. Journal of Usability Studies 1(4), 156–170 (2006)