Intercultural Usability Surveys: Do People Always Tell "The Truth"?

Emilie W. Gould

Adjunct Professor, Center for Distance Learning, SUNY Empire State College
111 West Avenue, Saratoga Springs, NY, USA 12866
goulde@alum.rpi.edu

Abstract. Researchers have identified many ways that culture affects usability methods – interviews, moderated tests, think-aloud protocols, and card sorts. This paper reviews some of that literature and discusses a project investigating the effect of culture on usability surveys.

Keywords: culture, cultural usability, survey methodology.

1 Introduction

Developing reliable, valid surveys for usability research is not easy. People may use words differently; apply different end-points (and middle-points) to scales; and situate answers in different social realities. By following principles of survey design, we can generally achieve statistical reliability. However, closer inspection of subsamples sometimes shows patterns of skipped questions and spoiled surveys – or a decreased or skewed range of answers.

Similarly, during pilot tests, interview guides for requirements gathering and usability reviews often change substantially. Researchers discover the questions that elicit abundant feedback and eliminate those that are redundant. Connotations emerge and words turn out to have different meanings in context.

However, we muddle on – assessing usability through processes that rely heavily on survey. We base our methods on assumptions that are part of the "culture of usability evaluation" [1].

Traditionally, we have assumed two things:

- we should ask all users the same questions to make it possible to compare feedback
- users will answer our questions truthfully

Within our own cultures, these assumptions may be more or less true. Unfortunately, when we begin to collect data in vastly different cultures, our assumptions are probably false.

2 What Kind of Methodology Problems Have People Already Found in Intercultural Evaluations?

Since the mid-1990s, people have been reporting problems with interviews, moderated tests, think-aloud protocols, and other methodologies used in international usability evaluations.

N. Aykin (Ed.): Internationalization, Design, LNCS 5623, pp. 254–258, 2009.

At CHI98, Alvin Yeo (then a student) discussed a usability study he had done to localize a spreadsheet. His evaluators were Malaysian staff at his university – higher and lower in status to himself. Despite bad experiences with the software, only those who were higher in status rated the software negatively; those with equal or lower status were more positive. Similarly, high-status evaluators made more negative comments and were "harsher" in the way they phrased them; low-status evaluators were more polite and "subtle." Yeo [2, 3] concluded that such self-censorship was based on the relationship of the evaluators to the test moderator. Malaysia is classified as a high power-distance, moderately high collectivist society by Hofstede [4]. In general, Yeo believed people sought to preserve harmony and save face – his own as well as their own – by refusing to be negative. By contrast, the high-status evaluator felt her problems had made her look incompetent so she criticized the software. Yeo suggested that Western usability assessment techniques should only be used with people who were already experienced users, familiar with the experimenter, and higher in status than the experimenter.

Needless to say, these recommendations pose some problems for our traditional focus on inexperienced users. In addition, they unbalance the experimental design of most usability assessments.

Apala Chavan [5] found similar "relational problems" in her research in India. Gender, youth, and class all affected the willingness of evaluators to talk about products. Women would often speak only with women; younger researchers had more success than older, more senior people. Unlike Yeo, Chavan attributed the difference to stronger social affiliations based on liking for people similar to one's self.

Clemmensen, Shi, Kumar, Li, Sun, and Yammiyavar [6] found support for both explanations in a study comparing the role of test moderators in usability assessments in India, China, and Denmark. Indian moderators had to deal with self-censorship, based on gender and age, among traditional end users in India. Male researchers needed to include a male relative in interviews with rural women, and older researchers "frightened" younger rural evaluators. In China, they found female moderators seemed to do better with male evaluators, an apparent reversal of power distance. However, in Denmark, usability tests ran most smoothly when the researchers and evaluators were the same gender, age, and shared the same level of job experience.

Clemmensen et al. suggested that intercultural usability tests need to:

- include "hidden user groups" (those less comfortable with foreigners or more traditional)
- estimate the "evaluator effect" and select researchers appropriate to those groups
- review the detection rate to see if these groups identify different types of problems
- modify the test protocols to localize scenarios, use more direct probes, or ask different questions

One suggestion for localizing test protocols adopts a technique developed by Chavan [5, 7]. She notes that Indian evaluators are often unwilling to criticize under any circumstance and recommends using dramatistic techniques rooted in Bollywood and traditional Indian theatre. For instance, Bollywood scenarios free people from the constraints of the "real world" and allow them to speculate about emotion and effects that belong to an idealized product or situation.

However, there could be some difficulty in applying this recommendation due to the aesthetic divergence between these traditions. Both began as sacred drama but the Western tradition focuses on using poetics (a counterpart to logic and rhetoric) to find concrete "truths," while the Indian tradition looks for release and transcendence.

> *The compulsions of the Indian theory of anukarana or imitation are different from the Greek ones. The success of anukarana is not judged in terms of its ability to represent the world but by its capacity to create a new world.*
>
> *The method of abhinaya … is not a mimesis of things but of bhavas (moods) which are ever changing in significance….*
>
> *One drama exploits free will, the other, destiny; one exploits tension, the other conspires to eliminate it.* (Chavan [7] quoting Gupt [8])

Although Chavan suggests art can be used as a medium to contextualize usability, there seem to be practical problems at the level of asking questions and interpreting answers.

Finally, usability researchers have examined the effect of culture on think-aloud protocols. Yeo [2] noted that most of his evaluators had difficulty sustaining a commentary on their actions. Clemmensen et al. [6] found that moderators running tests in China needed to use many more direct probes since evaluators would not identify their actions unless prompted. However, after a period of silence, many often provided a retrospective think-aloud analysis of their choices.

Shi [9], reviewing usability tests in Beijing, also found that Chinese people needed regular prompting. He attributed their silence to the holistic thinking style and interpersonal needs of East Asians. His explanation draws on Nisbett's [10] description of Asian and Western cognitive styles – the first derived from the five Confucian relationships and the second from Greek philosophy. Nisbett believes Asian thought focuses on relations among people and events, social harmony, and the acceptance of natural processes and change; Western thought is more attentive to objects, formal logic, categories, control, and stable theories of explanation. When faced with contradiction, Asian people tend to look for a middle way; Western people insist on correctness and "truth." Nawaz, Plocher, Clemmensen, Qu, and Sun [11] found support for holistic thinking among Chinese evaluators in a card-sorting exercise designed to test information structure.

All these studies – and there are many more now available in research journals and proceedings from the ACM, HCI International, and IWIPS – demonstrate that our usability methods are not as "methodical" as we once believed. Culture seems to affect interviews, moderated tests, think-aloud protocols, and card-sorts. Explanations can be found in theories of cultural dimensions, sociology, and cognitive differences. Recommendations include limiting (or expanding) types of evaluators, selecting test moderators and researchers on the basis of their similarity (or dissimilarity) to the evaluators, localizing scenarios, applying theories from drama to contextualize tests, and modifying probes.

Such diversity should not be unexpected; usability research began adopting qualitative methods in the 1980s and lack of generalizability is one of the key features of such research. Ethnographic methods provide rich data but that data must be "grounded"

and interpreted to make it useful for product development. We have begun to understand some of the ways culture affects tests and interviews. My current research looks at ways survey data may be skewed in intercultural usability assessments.

3 Does Bias Affect Surveys Used in Intercultural Research?

While working on my PhD thesis, I developed surveys for Malaysia and the United States [11]. Colleagues translated my terminology into Bahasa Malaysia but I am the first to recognize that some of the differences that I saw in my return rate may have reflected incomplete localization. However, I also found higher rates of deliberately spoiled surveys, plus skipped questions and skewed scales from Malaysian respondents. These seemed to reflect conscious choices based on attitudes, not accidents.

Such differences pose the following questions for research:

- Are there underlying cultural attitudes to survey research that lead to such behavior?
- Is the problem simply that surveys are less common in some countries?
- Do people in some countries actively resist data collection by survey? Does such resistance depend on who developed the survey or where the survey originated?
- Do people feel compelled to reveal everything in surveys? Is it ever okay to lie?
- Do survey scales match people's confidence in their answers?
- Do people prefer anonymity or situations where they have a relationship with the researcher?
- Are people more positive about participating in interviews than in surveys?
- How do these attitudes relate to multinational companies doing usability research?

I am currently collecting and analyzing data to answer these questions using focus groups with international students in the United States and Canada. Because such students have experienced two cultures, they are in a position to compare and contrast their experiences in dealing with surveys. However, this same experience also makes these students an atypical minority within their own countries. As a result, international colleagues are also conducting a limited number of focus groups to validate my results. Full results will be presented at HCII 2009.

References

1. Wallerstein, I.: The heritage of sociology, the promise of social science (1998), http://www.binghamton.edu/fbc/iwprad1.htm
2. Yeo, A.: Cultural effects in usability assessment. In: CHI 1998, pp. 74–75. ACM Press, New York (1998)
3. Yeo, A.: Are usability assessment techniques reliable in non-Western countries? Elec. J. on Info Sys in Dev Countries 3(1), 1–21 (2000)
4. Hofstede, G.: Cultures and organizations: Software of the mind. McGraw-Hill, New York (1997)
5. Chavan, A.: Usability in India, is it different? In: HCII 2005, Lawrence Erlbaum, Mahwah (2006)
6. Clemmensen, T., Shi, Q., Kumar, J., Li, H., Sun, X., Yammiyavar, P.: Cultural usability tests – how usability tests are not the same all over the world. In: Aykin, N. (ed.) HCII 2007. LNCS, vol. 4559, pp. 281–290. Springer, Heidelberg (2007)

7. Chavan, A.: What about a 'local' wrapper around an 'universal' core? In: CHI 2008, pp. 2605–2607. ACM Press, New York (2008)
8. Gupt, B.: Dramatic concepts, Greek and Indian: A study of Poetics and Natyasastra. D.K. Printworld, New Delhi (1994)
9. Shi, Q.: A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. In: NordiCHI 2008, pp. 344–352. ACM Press, New York (2008)
10. Nisbett, R.: The geography of thought: How Asians and Westerners think differently and why. Free Press, New York (2003)
11. Nawaz, A., Plocher, T., Clemmensen, T., Qu, W., Sun, X.: Cultural differences in the structure of categories in Denmark and China. Working paper 03-2007, Department of Informatics, Copenhagen Business School (2007)
12. Gould, E.W.: Applying cultural dimensions to website design: A case study from Malaysia and the United States. Doctoral dissertation, Rensselaer Polytechnic Institute, 2004. In: Dissertation Abstracts International (2004)