ICIS Proceedings

ICIS 2008 Proceedings

Association for Information Systems *Year* 2008

Cultural Cognition in the Thinking-Aloud Method for Usability Evaluation

Torkil Clemmensen[*] Morten Hertzum[†] Kasper Hornbaek[‡]
Qingzin Shi[**] Pradeep Yammiyavar[††]

[*]Copenhagen Business School, tc.inf@cbs.dk
[†]Roskilde University, mhz@ruc.dk
[‡]University of Copenhangen, kash@diku.dk
[**]Copenhagen Business School, qs.inf@cbs.dk
[††]India Institute of Technology, pradeep@iitg.ernet.in

CULTURAL COGNITION IN THE THINKING-ALOUD METHOD FOR USABILITY EVALUATION

Cognition culturelle pour l'évaluation de l'usabilité par la méthode de pensée à haute voix

Completed Research Paper

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

Morten Hertzum
Computer Science,
Roskilde University,
Roskilde, Denmark
mhz@ruc.dk

Kasper Hornbæk
Department of Computer Science,
University of Copenhagen
Copenhagen, Denmark
kash@diku.dk

Qingxin Shi
Department of Informatics,
Copenhagen Business School
Frederiksberg, Denmark
qs.inf@cbs.dk

Pradeep Yammiyavar
Department of Design,
Indian Institute of Technology,
Guwahati, Assam, India
pradeep@iitg.ernet.in

Abstract

We discuss the impact of cultural differences on usability evaluations that are based on the thinking-aloud method (TA). The term 'cultural differences' helps distinguish differences in the perception and thinking of Westerners (people from Western Europe and US citizens with European origins) and Easterners (people from China and the countries heavily influenced by its culture). We illustrate the impact of cultural cognition on four central elements of TA: (1) instructions and tasks, (2) the user's verbalizations, (3) the evaluator's reading of the user, and (4) the overall relationship between user and evaluator. In conclusion, we point to the importance of matching the task presentation to users' cultural background, the different effects of thinking aloud on task performance between Easterners and Westerners, the differences in nonverbal behaviour that affect usability problem detection, and, finally, the complexity of the overall relationship between user and evaluator when they have different cultural backgrounds.

Keywords: cultural differences, thinking aloud, usability, international systems development

Résumé

Nous discutons de l'impact des différences culturelles sur les évaluations d'usabilité basées sur les méthodes de pensée à haute voix. Nous soulignons : l'importance de faire correspondre la présentation des tâches au bagage culturel de l'utilisateur, les différents effets de la pensée à voix haute sur la performance des tâches entre orientaux et occidentaux, les différences de comportement non verbaux qui affectent la détection des problèmes d'usabibité, et la complexité d'une relation utilisateur-évaluateur interculturelle.

摘要

我们讨论了文化差异对可用性评价的影响，以出声思维法为例进行了分析。本文阐述了任务表述方式和用户文化背景相匹配的重要性，讨论了出声思维对东方人和西方人任务绩效的不同影响，以及非言语行为对发现可用性问题的不同作用，并对不同文化间的用户-评估者关系的复杂性进行了论述。

Introduction

Information technology is being developed and used by diverse groups of people. Reasons for this include the outsourcing of development activities, for example, from Europe to Asia, the emergence of globally available systems such as mobile phones and web-based applications, and the increasing use of localized information technology by local-language populations across the world. In attempting to disentangle this diversity, culture has received increasing attention in software engineering (Sahay et al., 2003), human-computer interaction (HCI; del Galdo & Nielsen, 1996; Smith & Yetim, 2004), and usability evaluation (Murphy, 2001; Yeo, 2001). However, most research on usability evaluation methods presupposes that usability evaluation is unaffected by cultural issues. For example, the cultural background of experimental participants is rarely reported, task scenarios are assumed to be culturally unbiased, interface heuristics are presented as universals, and disagreements between studies are rarely discussed in terms of cultural effects. Given that definitions of usability (e.g., ISO 9241, 1998) and HCI in general assign importance to concepts such as the context of use, which includes users' cultural background, this is surprising.

This study examines differences in cultural cognition in the thinking-aloud method. The presented arguments should be valid for the thinking-aloud method in general, but the context that is used to illustrate the arguments is usability evaluation. Usability evaluation based on the thinking-aloud method (henceforth, TA) is in widespread use and commonly considered the single most valuable method for usability evaluation (Clemmensen, 2005; Gulliksen et al., 2004; Nielsen, 1993). In essence, TA consists of a user that thinks out loud while using a system, and an evaluator that observes the user and listens in on his or her thoughts. The research on TA is considerable and includes studies of the number of users needed for finding a specified proportion of the usability problems in an application (Lewis, 1994), the potential of having users work in pairs (van den Haak et al., 2004), the evaluator's effect on the set of problems identified (Hertzum & Jacobsen, 2003), and the validity of verbalization (Boren & Ramey, 2000). Recent work on the influence of culture on people's perception and thinking yields results that call into question whether our knowledge and assumptions about such basic characteristics of TA are valid outside of Europe and the US. The aim of this study is to introduce the cultural psychology of Nisbett (2003) as a conceptual basis for thinking about TA and to analyse the influence of culture on TA.

Culture is a complex concept. In this study, we use the term culture as a means of distinguishing among differences in cognitive style, i.e. differences in the perception and thinking, of people with a background from majority cultures in different regions of the world. Following Nisbett (2003), we primarily focus on cultural differences between Westerners (people from Western Europe and US citizens with European origins) and Easterners (people from China and the countries heavily influenced by its culture, such as Korea). These groups show patent cultural differences, but Hofstede (2001) makes it evident that cultural differences exist between numerous countries, not just between Westerners and Easterners. For example, Nisbett's work may also be valid and important for usability testing in India.

This paper is not an analysis of cultural differences in what makes a good interface (for a review, see Callahan, 2005). Rather, we seek to provide a conceptual analysis of the influence of culture on various aspects of performing usability evaluations based on TA.

In the next section we introduce Nisbett's (2003) work on cultural cognition. Then, we provide a simplified overview of TA as a reference model for the subsequent analysis. Our analysis in the subsequent section on the influence of culture on TA consists of applying the work on cultural cognition to four central elements of TA: (1) instructions and tasks, (2) the user's verbalizations, (3) the evaluator's reading of the user, and (4) the overall relationship between user and evaluator. In conclusion, we discuss implications for practical usability evaluation as well as for usability research.

Cultural cognition

Basic human psychological characteristics are often seen as universal, implying that people across the world perceive and reason in the same way (Brown, 1991; Pinker, 2006). Similarly, software engineers and usability specialists probably tend to assume that when a Chinese user and a European user look at the same web page, they perceive the same web page – though they may interpret the information on it in different ways. Nisbett (2003; see also Nisbett et al., 2001) provides compelling evidence against such universalism and argues that cultural-historical differences in physical environment, upbringing, education, and social structure shape how people from different regions of the world perceive objects and situations.

Nisbett (2003) focuses in particular on two broad groups of people – Easterners and Westerners – because they inhabit two regions of the world that since ancient time have had different intellectual traditions and thereby different cultural bases for cognition. Specifically, Easterners are primarily the people of China, Korea, and Japan; Westerners are primarily Europeans and Americans. Nisbett finds that Westerners' way of thinking can be characterized as analytic – they tend to "think in a line" – whereas Easterners' way of thinking is more holistic – they tend to "think in a circle". This overall characterization is based on a series of experiments, which demonstrate consistent differences between Easterners and Westerners in several main aspects of cognition.

One example of such differences is *what people attend to*. According to Nisbett (2003), Easterners attend more to environments and Westerners more to objects. Furthermore, Easterners are more likely to detect relationships among events than are Westerners. An illustration of these differences is provided by studies of field dependence, a psychological construct that describes the extent to which a person's perception of an object is influenced by the environment in which the object is set (Witkin et al., 1977). In one such study Masuda and Nisbett (2001) showed Japanese and American participants animations of fish swimming in settings of plants, rocks, and other underwater objects. When asked to report the contents of the animations, American and Japanese participants made equally many references to the focal fish, but Japanese participants made more statements about the settings and relationships than American participants. In a recognition test, Japanese participants recognized previously seen objects more accurately when they saw them in their original settings rather than in novel settings; this manipulation had little effect on American participants. Other studies of field dependence report similar differences (Ji et al., 2000; Kühnen et al., 2001a) and suggest that field dependence can to some extent be induced through priming (Kühnen et al., 2001b).

Another difference in cognition concerns *the extent to which people experience surprise*. According to Nisbett (2003), Easterners expect frequent, dynamic change and therefore experience less surprise in face of the manifold ways in which things evolve, compared to Westerners who to a larger extent perceive their world by means of logic and therefore notice – with surprise – when things evolve in inconsistent ways. As one example, Choi and Nisbett (2000) presented Americans and Koreans with either two conflicting hypotheses about a research study or with just one hypothesis that predicted the outcome of the study. In the case of two hypotheses, one of them predicted the outcome of the study. Upon reading a description of the study Americans found it more surprising when they had read the two hypotheses; Koreans showed no difference in surprise from when they had only read one hypothesis. In explaining the difference in presence or absence of surprise, Nisbett (2003) refers to studies showing that Westerners are more likely to attribute a course of events to a few causes or actors, whereas Easterners consider more factors and emphasize the context (Miller, 1984). Considering a large number of factors appears to make it easier to think of a reason why a particular event turned out the way it did, leading to less surprise.

People also *group objects in culture-dependent ways*. According to Nisbett (2003), one clear difference is whether people group according to taxonomic categories or thematic relationships. For instance, Ji et al. (2004) asked students to indicate which two out of three words (e.g., panda, monkey, and banana) were most closely related. American students tended to group on taxonomic category (i.e., panda and monkey) while Chinese students tended to group on thematic relationship (i.e., monkey and banana). Irrespective of whether they were presented with words in English or Chinese, bilingual Chinese students grouped words in a more relational and less categorical way than American students, indicating that the difference in grouping is not merely an effect of language. Rather, Westerners seem to organize their knowledge in terms of categories, while Easterners organize theirs in terms of relationships.

Finally, Westerners and Easterners appear to *deal differently with seeming contradictions*. According to Nisbett (2003), when presented with evidence of apparently contradictory propositions Westerners are inclined to reject one proposition in favour of the other, while Easterners tend to try to find truth in both propositions. Peng and Nisbett (1999) presented students with either single research findings or pairs of research findings that seemed to contradict each other. Based on the students' rating of the plausibility of the research findings when they were seen without contradictory information, the more plausible and the less plausible research findings were identified. American students presented with pairs of contradictory research findings rated the more plausible findings as even more plausible than if they saw no contradictory information; there was no difference in their ratings of the less plausible research findings. In contrast, Chinese students presented with pairs of contradictory research findings rated the less plausible research findings as more plausible than if they saw no contradictory information. That is, American students seemed to follow a polarization approach seeking for the one right way, whereas Chinese students seemed to follow a compromise approach seeking for a middle way.

Nisbett's results show that culture affects individual experiences on a very basic perceptual level. It appears, however, that two limitations must be kept in mind in discussing Nisbett's work. First, it is based on studies of college and university students. Students are generally young, well educated, and unlikely to be representative of the entire population. It is unclear how this may affect the results. Second, some studies suggest that analytic and holistic thinking can be induced by dynamic, situational factors (Briley et al., 2000; Hong & Mallorie, 2004). Thus, analytic and holistic thinking may either be intrinsic to participants' cultural background or be two cognitive styles that are available across cultures and may be triggered by situational factors or even used knowingly in situation-dependent ways.

A simplified model of TA

After TA was introduced as a usability evaluation method, numerous variations of the method have been employed. Today there is no definitive definition of the method and no single accepted procedure for usability specialists to follow (Dumas, 2003; Lewis, 2006). TA is normally performed by having individual users think out loud while interacting with a system. It can also be performed with pairs of cooperating users (van den Haak et al., 2004) and with users that verbalize retrospectively rather than concurrently (Page & Rahimi, 1995). Further, the interaction between user and evaluator may be restricted to users' verbalizations of their thoughts and occasional reminders to "keep talking" (in line with Ericsson & Simon, 1993) or it may include requests for elaboration and reflection (e.g., Monk et al., 1993). Finally, practitioners' ways of actually performing TA do not always follow methodological prescriptions (Boren & Ramey, 2000; Nørgaard & Hornbæk, 2006).

Our goal is not to establish a consensus about how to conduct TA, but simply to point out the main elements of the method. Across the numerous variations of TA these main elements recur. Our simplified model of TA is presented in Figure 1. It comprises four main elements, which structure the subsequent analysis of the extent to which TA is susceptible to cultural effects:

- *Instructions and tasks*. Users interact with the system based on a set of instructions, including an explanation of how to think aloud, and a set of tasks prepared ahead of the evaluation.
- *Verbalization*. While solving the tasks users verbalize their thoughts. If users fall silent for longer periods of time they are prompted to resume verbalization. These prompts may vary from a detached "Keep talking" to questions asking users to explain their behaviour.

- *Reading the user*. One or more evaluators observe the users' behaviour and listen in on their thoughts. On this basis evaluators extract, and subsequently describe and report, usability problems.
- *Overall relationship between user and evaluator*. While the evaluator attempts to establish a situation in which the user feels free to make both positive and negative comments, the user may not feel at ease. A productive relationship hinges on issues such as language and indirect communication cues.

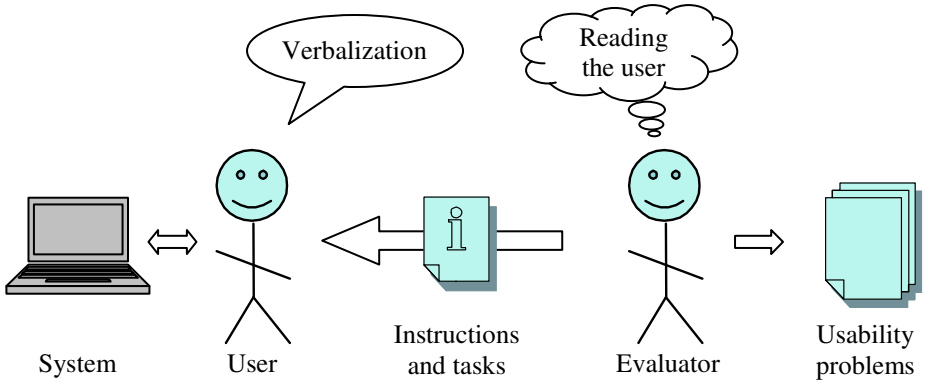

Figure 1. Reference model of TA.

TA is a tool borrowed from Western cognitive science (Lewis, 1982), and it has in recent years spread rapidly also in East Asian countries. This makes Nisbett's (2003) work pertinent in relation to the practice of TA as well as to research on TA. If TA implicitly assumes analytic thinking, it needs to be clarified how this affects usability evaluations in which users, evaluators, or both are holistic in their style of thought.

The influence of culture on TA

In what follows, we discuss how culture is likely to affect important characteristics of TA, and contrast Nisbett's findings with conventional thinking about TA. The focus is solely on the TA test session, disregarding issues such as planning and reporting of the TA.

Instructions and tasks

The tasks are an important element of TA and may introduce various biases (Cordes, 2001). In presenting tasks, evaluators generally choose between two kinds of presentation. One way is to merely instruct users what to do (e.g., "Please, save the phone number 6496 7721 on the mobile phone being evaluated") as explained by, for example, Preece et al. (2007). The other way, recommended by, for example, Dumas and Redish (1999), is to embed the instructions in a scenario (e.g., "You've just got a new mobile phone. The number of your best friend, Chris, is 6496 7721. Please, save Chris' number on your new mobile phone."). The two kinds of presentation differ with respect to whether they provide contextual information in addition to the focal information and may thus, according to Nisbett, be perceived differently by Easterners and Westerners because Easterners tend to be more field dependent than Westerners.

The two kinds of presentation may appear quite similar to Westerners, who primarily attend to the focal information, which is present in both kinds, and pay less attention to the contextual information, which may be seen as superfluous. Eastern users may, however, perceive the two kinds of presentation differently because Easterners attend to contextual information when it is present and may find context-free presentation of focal information artificial and hard to understand. Support for such cultural differences is provided by Miyamoto et al. (2006), who found that Westerners primarily attended to the focal part of a picture and separated objects from their environment, whereas Easterners attended more to contextual information and did not separate objects from their environment. Similar differences hold for social situations. For example, Chua et al. (2005) found that when American and Chinese students were asked to read descriptions of social events, American students tended to focus on main

characters and perceived actions as the results of these characters' intentions, while Chinese students also attended to peripheral characters and ascribed actions to circumstances. In the context of website development, Smith et al. (2004) concluded that task scenarios must be richer for effective use with Indian users.

Preece et al. (2007, p. 654) gave an example of a procedure for user testing in which the tasks are introduced with the statement "As you work on each task, I'd like you to imagine that it's something you or someone close to you needs to know." While this suggests a scenario approach, the introduction is followed by a list of apparently unrelated tasks: "Task 1: find information about whether a dark lump on your shoulder might be skin cancer. Task 2: find information about whether it's safe to use Prozac during pregnancy…" Such a disguised way of presenting tasks by mere instructions is likely to be particularly confusing to Easterners because they will be attentive to the overall framing suggested by the introductory statement, but find it unused in the sequence of tasks. According to Witkin et al. (1977) field-dependent persons have a greater need for a clear structuring of tasks in order to assign prominence to relevant cues and avoid confusion. Conversely, field-independent persons are affected less by unclear relations between contextual and focal information because they pay less attention to contextual information.

The Bollywood method (Chavan, 2005), a variant of the scenario approach, aims at making Indian users disclose their thoughts more openly during TA sessions. The method has users imagine a dramatic scenario similar to those most Indians are used to from Bollywood movies. Reviews of Bollywood movies are one of the few popular and accepted forms of openly expressing critique in India, and the Bollywood method portrays usability evaluation in terms of this movie-review genre, hoping that users will transfer the open critiquing behaviour to the usability evaluation. Thus, the method builds on people's tendency to mimic the behaviour associated with their current environment. Field-dependent persons are more likely to mimic other people's behaviour than are field-independent people (van Baaren et al., 2004). Thus, the Bollywood method may be particularly effective for Easterners, especially Indians because they are presented with a culturally established form of expression.

Tasks also serve as a vehicle for the evaluator to keep track of what the user is trying to do. To accurately provide such information, the tasks must be perceived similarly by user and evaluator. In relation to TA, Smagorinsky (1998) has challenged the assumption that tasks have such shared or socially agreed upon performance characteristics. Smagorinsky found that across cultural settings users and evaluators may not converge on a shared representation of tasks and that the divergences in their task representations may suggest different paths to task completion. As an example, a Western evaluator might not see the task of writing an invitation as that of designing both the text and the paper, and would therefore not share an Eastern user's pre-occupation with the look and colour of the paper background in a word processor. Thus, in addition to being perceived differently by users with different cultural backgrounds tasks may be perceived differently by users and evaluators with different cultural backgrounds, complicating the analysis of users' behaviour.

The user's verbalization

Compared to other methods of usability evaluation, TA gives evaluators access to otherwise inaccessible information about what users are thinking (Nielsen et al., 2002). However, this advantage of TA rests on two premises. First, asking users to think aloud should not affect their performance. Second, users' verbalizations should be valid expressions of their thoughts. If one of these premises does not hold, TA will not accurately reflect real use of the evaluated system and problems identified during TA sessions might not be indicative of the problems users will experience during real use of the system.

The theoretical underpinning for TA is Ericsson and Simon's (1980; 1993) classic model of verbalization. According to Ericsson and Simon, TA gives a valid expression of users' thoughts and leaves task performance unaffected as long as users verbalize information in their current focus of attention (verbalization at levels 1 and 2) and refrain from providing explanations and retrieving additional information from memory (verbalization at level 3). This is assumed to hold irrespective of culture. Contrary to this assumption, Nisbett (2003, p. 211) sees a fundamental difference in how easily Westerners and Easterners may express their thoughts:

Analytic thought, which dissects the world into a limited number of discrete objects having particular attributes that can be categorized in clear ways, lends itself to being captured in language. Holistic thought, which responds to a much wider array of objects and their relations, and which makes fewer sharp distinctions among attributes or categories, is less well suited to linguistic representation.

On this basis we propose that whereas TA may be sufficiently easy for Westerners to not interfere with their task performance, TA is so difficult for Easterners that it impairs their task performance.

Kim (2002) had Westerners and Easterners solve reasoning problems while they were thinking out loud. All participants were university students in the US, but the Westerners were third or older generation Americans (i.e., both of their parents were also born and raised in the US), whereas the Easterners were second-generation Americans (i.e., both of their parents were immigrants from East Asian countries). In Study 1 participants were divided into a control group that simply received instructions about the tasks and a thinking-aloud group that was also instructed to think aloud while solving the tasks. In Study 2 participants first solved a series of tasks in silence and then solved a series of tasks while thinking aloud. In both studies the thinking-aloud instructions complied with Ericsson and Simon's (1993) prescriptions for verbalization at levels 1 and 2. As predicted by Ericsson and Simon's model the number of tasks solved correctly by Westerners was not impaired by thinking aloud. The results of Study 2 even showed the opposite effect, namely that forcing Westerners to perform in silence significantly impaired their performance. This suggests that Westerners may habitually use talking as a means of supporting their thinking. In contrast, Easterners' performance was significantly impaired by thinking aloud in both Studies 1 and 2. In Study 1, Easterners gave incorrect answers to about 34% more tasks when thinking aloud (9.24 correct of 20 tasks), compared to when they performed without thinking aloud (12.35 correct of 20 tasks). Thus, thinking aloud appears to be foreign to Easterners, to the extent that their task performance is degraded significantly.

The tendency of Westerners to perform better when they think aloud was further explored in Study 3 (Kim, 2002). There participants had to either solve 10 tasks while staying silent and then 10 tasks while verbalizing, or first stay silent and then solve 10 tasks while repeating the alphabet out loud. As hypothesized, European Americans improved their performance while thinking aloud compared to performing in silence (from 4.85 to 5.30 correct tasks) and were greatly hindered by having to do another verbal task (saying the alphabet out loud) compared to performing in silence (5.05 versus 3.73 correct tasks). Conversely, Asian Americans were only hindered by the thinking-aloud condition (6.18 versus 4.73 correct tasks) and not by saying the alphabet out loud. One explanation for these results is that Westerners may benefit from thinking aloud because it accords with processes they habitually use to support their thinking, while Easterners to a lesser extent rely on verbal representations in supporting their thought processes.

In practical TA sessions, thinking aloud is often not restricted to verbalization at levels 1 and 2, during which the person thinking out loud suspends any awareness of listeners, but also includes elements of conversation and interview (Boren & Ramey, 2000). In addition to general questions about the validity of the obtained data, this may further allow cultural differences to influence TA. Briley et al. (2000) found that asking Easterners to provide reasons for their choices changed their behaviour compared to when they were not asked for reasons. When asked for reasons for their choices, Easterners became more likely to choose the middle option, which offered moderate levels of two attributes, as opposed to one-attribute options, which offered either one or the other attribute. Conversely, Westerners became more likely to choose a one-attribute option when they were asked to verbalize their reasons. Briley et al. (2000) drew two conclusions from their results. First, the studies are consistent with participants' culture in that Western, analytic thought favours one-attribute options whereas Eastern, holistic thought favours the middle option. Second, culture is brought into the choice situation only when somehow evoked, for example by asking participants to provide reasons. Asking participants to provide reasons shifts their focus from a search for the best option to a search for the option supported by the best reasons. Otherwise participants' choices would not be affected by whether or not they were asked to provide reasons for their choices. Thus, asking TA participants to provide reasons for their behaviour is likely to affect their performance in culture-dependent ways.

With respect to whether users' verbalizations can be taken as valid expressions of their thoughts, Nisbett (2003) suggests that the language in which verbalizations are made affects their content. According to Nisbett (2003, p. 155), Westerners tend to speak in context-free sentences, use words with distinctive meaning, emphasize agents and other focal objects, and use markers to identify whether they speak about a category (e.g., "buttons are round") or an instance (e.g., "this button is round"). Conversely, Easterners rely on sentences being understood in the context of surrounding sentences, emphasize relations and context, and use the speech context, not language markers, to indicate whether they talk about a category or an instance. Furthermore, studies of TA show that recall of stories depends partly on language characteristics. Haritos and Nelson (2001) found that bilingual children recalled more of a story when asked to recall in Greek a story presented in English than children asked to recall the same story in

English. In addition, Evers (2002) found that verbalization is easier for North Americans than for Japanese users who feel uncomfortable verbalizing their thoughts.

Reading the user

A main activity in TA is that evaluators listen to users' verbalizations and observe their facial expressions and gestures in order to report the usability problems that users experience. The literature on usability evaluation suggests that usability evaluators involved in the same test, say observers in TA, report different usability problems (e.g., Hertzum & Jacobsen, 2003; Vermeeren et al., 2003). Hertzum and Jacobsen (2003, p. 201) suggested that an important reason behind the difference in usability-problem reports is that "usability evaluation is a cognitive activity, which requires that the evaluators exercise judgment". Given the data on cultural differences presented by Nisbett (2003) and others, it may be expected that these judgements are to some extent influenced by the evaluator's cultural background, leading to the reporting of different usability problems. Below, we first discuss how the activity of reading the user may differ depending on the evaluators' cultural background and then turn to some specific cultural biases in generating and grouping usability problems from evaluators' observations.

As the section on cultural cognition points out, Nisbett (2003) argues that a major difference between Westerners and Easterners concerns how observations are made, in particular what objects and actions persons with different cultural background attend to and which observations they consider surprising. One mechanism behind this difference is field dependence, which in a TA context concerns the extent to which an evaluator's perception of difficulties and usability problems is influenced by the environment/context in which they occur. We hypothesize that the number of problems evaluators attend to differs depending on evaluators' field dependence. Thus, more problems might be reported by field-independent evaluators because they rely less on environmental cues (such as the user's gestures or facial expressions) as input for rejecting candidate usability problems. It also seems that different kinds of usability problem might be detected by field-dependent and field-independent evaluators because they – at least on some occasions – attend to different aspects of a user's experience with the user interface. In a study of heuristic evaluation, Ling (2005) documented how field-independent evaluators made more frequent use of analytic approaches when conducting their evaluation compared to field-dependent evaluators. Similar differences may be expected in evaluators' approaches to reading the users in TA.

With respect to reading the user, a main difference relevant to TA concerns the extent to which Easterners and Westerners experience surprise. As a consequence different evaluators may find users' behaviour more or less surprising and therefore differ in their judgement as to whether it indicates a usability problem. In a study of Western evaluators, Nørgaard and Hornbæk (2006) argued that evaluators in TA may be focusing too much on already known usability problems. The cultural difference in experience of surprise suggests that this tendency to overlook the unexpected may be even stronger for Eastern evaluators. Moreover, many usability studies employ users' spontaneous or prompted expression of surprise as a way of identifying usability problems. Looking for such expression of surprise has been suggested as valuable indicators of when to consider something a problem (Jacobsen et al., 1998) or when to probe users for explanations so as to identify problems (Rubin, 1994).

In addition to sheer differences in reading the user during a TA session, evaluators' cultural background may affect how observations of the user are analysed and turned into usability reports. An important part of describing usability problems is the analysis of causes (Lavery et al., 1997). The data on cultural differences presented by Nisbett (2003) suggest that the attribution of causes differs across cultures. This could mean that evaluators from different cultures emphasize different events and user-interface elements when analysing user behaviour and thus report different usability problems. Nisbett (2003, p. 114) notes, for example, that "[…] Chinese people are inclined to attribute behaviour to context and Americans tend to attribute the same behaviour to the actor". Further, it seems that Westerners are more disposed to embark on causal attribution; that is, to explain observed phenomena as caused by a few factors.

If these effects hold also for usability evaluators, it may have a number of implications. Most importantly evaluators may differ in the extent to which they clearly identify a factor behind some observed difficulty; the number of problems that are attributed to users' personality traits (e.g., being slow and inexperienced) may also differ. Related to this discussion is Choi and Nisbett's (1998) study of the correspondence bias; that is, the assumption that an individual's behaviour accurately reflects that individual's dispositions. They found that Korean subjects were less susceptible to the correspondence bias than American subjects when judging a person's attitude based on an essay

the person had written with or without a choice as to the content of the essay. Thus, Koreans seemed more attentive to situational factors in forming their judgement about a person's attitude. In relation to TA this suggests that Western evaluators may generate simpler explanations for problems and possibly disregard some problems because the evaluators attribute them to user dispositions or traits, rather than to situational factors (e.g., the tasks and the setup of the evaluation).

Another part of the analysis of TA sessions consists of grouping observations across tasks and users. For instance, one of the criteria for treating something as a usability problem put forward by Jacobsen et al. (1998) is that "the evaluator generalizes a group of previously detected problems into a new problem". Here we may expect differences between cultures because there may be different rationales for generalizations. As mentioned in the section on cultural cognition, Nisbett (2003) finds that while Easterners group objects by thematic relationship, Westerners group objects by taxonomic category. It should therefore be expected that evaluators with different cultural backgrounds will group usability problems differently, leading to differences in the sets of usability problems reported after a series of TA sessions.

The overall relationship between user and evaluator

The relationship between the evaluator and the user in a TA session is shaping what users say when thinking aloud, how well evaluators read the user, whether the situation is experienced as artificial or natural, and much more. Yet, it is often assumed that the user's perception of the evaluator does not seriously affect the results of TA. Ericsson and Simon (1993) prescribed that to maintain valid verbalization (i.e., verbalization at levels 1 and 2) the evaluator should tell the user to speak "as if alone in the room". This approach minimizes interaction and communication between user and evaluator. In practice, much more interaction seems to take place in TA sessions, meaning that "talk is not simply a form of action" performed by the user alone, "but a mode of interaction" between user and evaluator (Boren & Ramey, 2000). Thus, the overall relationship between user and evaluator may be affected by a host of psychological, social, cultural, and other contextual factors. Here we focus on the role of culture.

When the evaluator and the user have different cultural backgrounds they cannot rely on a shared repertoire of cultural habits and manners of speech, such as a shared understanding of irony, metaphors, and under- and overstatements. This may decrease the effectiveness of the communication. For example, Vatrapu and Pérez-Quiñones (2006) conducted usability-evaluation interviews with 16 Indian participants, half run by an Indian evaluator and half by an Anglo-American evaluator. They found that when evaluator and user had similar cultural backgrounds more usability problems were found and more suggestions were made by users. One explanation of this finding could relate to conversational indirectness, which denotes the extent to which the literal meaning of an utterance relates to what the speaker intends to communicate. Compared to Westerners, Easterners have been found to rely more on conversational indirectness, particularly in work-related communication (Sanchez-Burks et al., 2003). As a common instance of this difference, Nisbett (2003) points to the difficulty amongst Westerners to read Easterners because Easterners are likely to assume that their point has been made indirectly and with finesse. Conversely, Easterners are apt to find Westerners direct to the point of condescension (Nisbett 2003, p. 61). These issues severely complicate efforts to apply TA in cross-cultural settings. Easterners are less confrontational when expressing themselves and this could lead a Western evaluator to conclude that they experience a higher level of satisfaction than Western users. Reversing the argument, an Eastern evaluator experiencing the direct non-conformist expressions of Western users may conclude that the usability of the system is worse than actually experienced by the user.

Furthermore, the communication between evaluator and user during a TA session may be affected by whether social relations are undertaken with a task focus or a socio-emotional focus (Sanchez-Burks et al., 2000). With a task focus people's effort is directed toward task-related goals, and their attention is focused on monitoring the extent to which these goals are being accomplished. With a socio-emotional focus, people's effort and attention are directed toward the interpersonal climate of the situation, and they strive to maintain social harmony. The Western culture is typically task-focused, which means that Western users' perception of the evaluator may not influence their behaviour appreciably, because they just focus on their task. In contrast, Easterners have a socio-emotional orientation, so Eastern users may be influenced more by their perception of the evaluator's status, background, and other characteristics. For example, Yeo (2001) showed that if Malaysian users in a TA session were not familiar with the evaluator, they made positive comments about a system even when they performed poorly. The users were

not only attending to the task but also trying to establish a harmonic relationship with the evaluator by avoiding too negative comments. Conversely, if the Malaysian users were familiar with the evaluator, they would not mind to make negative comments, because a socio-emotional relationship with the evaluator was already in place and they were confident their negative comments would not damage this relationship.

Finally, the language spoken during a TA session may affect the outcome. Ji et al. (2004) found that for bilingual Chinese who grew up in an environment dominated by one language and culture and only later learned the other language and culture, it mattered whether the test language was Chinese or English. When tested in Chinese the participants' responses were more in line with a Chinese mode of thinking than when they were tested in English. For bilingual Chinese who grew up in a mixed linguistic and cultural environment, the test language had no influence on their responses. This indicates that for some users the language in which an evaluation is conducted may prime their mode of thinking – testing in Chinese may foster a holistic mode of thinking, while testing in English may foster an analytic mode of thinking. Chinese tested in English were still more holistic in their mode of thinking than European Americans, but the language manipulation significantly affected the magnitude of the difference (Ji et al., 2004). Thus, priming by virtue of choosing a particular test language may enhance or diminish the cultural effects seen in TA, depending on whether users speak the language of their culture or of another culture. Moreover, the user may sometimes temporarily switch language in the midst of a TA session to compensate for lack of mastery of one language or to communicate what can best be communicated in another language. Evaluators may choose to demonstrate thinking aloud to bilingual users in their native language and to use native-language task descriptions and test materials in order to avoid priming effects and frequent switching of language (Wang & Wen, 2002). Alternatively, evaluators may deliberately mix languages in their instructions to demonstrate to users that they may use more than one language for verbalizing (Wang, 2003).

Discussion

Our analysis of the TA method by use of Nisbett's cultural psychology suggests that culture influences how instructions are acted upon by users, how users verbalize, how evaluators read users, and how the overall relationship between evaluators and users develops. These influences have implications for practitioners and for researchers, in particular those wishing to do cross-cultural work. Below we discuss these implications.

Advice for practitioners doing TA tests

Table 1 summarizes our advice for practitioners who do TA tests, in particular for Western evaluators intending to test in Eastern cultural contexts. Compared to existing guidelines on international use of TA tests (e.g., del Galdo & Nielsen, 1996), our advice is grounded in principles of cultural psychology. Some of the advice is related to common recommendations for how to do TA tests, for instance those of Rubin (1994), but again the advice differs by providing a psychological motivation.

With respect to tasks and instructions, practitioners should consider the difference in how Easterners and Westerners attend to objects and environments. It should be ensured that the background of the test, its goal, its tasks, and any instructions are made clear to users, especially if they are Easterners. Such users are more likely to want to know this background information and they are more likely to act on it, compared to the typical Western user. Also, practitioners should ensure coherence between the framing of tasks (e.g., in a scenario) and the actual tasks to be carried out.

With respect to the users' verbalization, practitioners should be aware that culture affects verbalization in multiple ways. For instance, it is more difficult to verbalize holistic thinking than analytic thinking, implying that Easterners will speak less during a TA session. It might be necessary to provide the user with more training in thinking aloud and to give prompts to think aloud more frequently than would normally be done. Note that Easterners may have poorer task completion rates when thinking aloud as opposed to when they are not thinking aloud. For Western users the opposite should be expected, especially for difficult tasks.

With respect to reading the user, practitioners should not base their evaluations on an assessment of users' expressions of surprise if they test with Easterners. Both evaluators and users may be affected by their cultural background when they attribute causes to problems. Western evaluators may attribute problems to the user's dispositions or traits (e.g., being slow and inexperienced), rather than to situational factors (e.g., tasks and evaluation

setup). Eastern users may not see the problem this way. Conversely, Western users may more frequently attribute difficulties to their dispositions or traits. In analysing and synthesizing readings of users, practitioners should be aware that evaluators with different cultural backgrounds might group usability problems differently, leading to differences in the sets of usability problems reported after a series of TA sessions.

With respect to the overall relationship between user and evaluator, several subtle effects are at play. As a result of conversational indirectness, expressions from Eastern users may seem vague or unclear with respect to preference. Practitioners should not conclude that an Eastern user is satisfied just because no open critique is voiced. Conversely, when testing with Westerners more accentuated opinions should be expected as should fewer attempts to find a middle way. Practitioners should not just focus on the test and what needs to be done, but take into account that Eastern users will be oriented toward the development of a harmonic socio-emotional relationship with the evaluator.

Table 1. Advice for practical thinking aloud tests in cross-cultural settings; in particular for Western evaluators with Eastern users.

Advice	Explanation and corresponding psychological principle
Explain the background of the test	Easterners want to know the broader context and background of a test; Westerners are less likely to focus on it. This is related to differences in what people attend to (Sections 'Cultural cognition' and 'Instructions and tasks').
Allow for more pauses when Easterners think aloud	Easterners have more difficulty in thinking aloud (Section 'The user's verbalization').
Thinking aloud might adversely affect Easterners' task performance	Thinking aloud might impair the performance of Easterners and enhance the performance of Westerners. Relates to the principles about expressing holistic and analytic thinking verbally (Section 'The user's verbalization').
Rely less on expressions of surprise when Easterners are test participants	The extent to which people express surprise differs between cultures (Section 'Reading the user'). Using surprise as a main marker of usability problems is thus problematic.
Be aware of and mitigate cross-cultural biases in analysing TA results	The attribution of causes to behaviour differs among cultures (Sections 'Reading the user' and 'The overall relationship between user and evaluator'). Further, the grouping and perception of similarities among behaviours and usability problems may differ depending on the evaluators' cultural background.
Critique of interfaces is likely to seek a compromise and be indirect when users are Easterners	Easterners use conversational indirectness and often attempt to find a middle path (Section 'The overall relationship between user and evaluator').
Use evaluators and users with similar cultural backgrounds, if possible	Difference in culture may impact the number of identified problems and redesign proposals. Familiarity between evaluator and user may also impact results (Section 'The overall relationship between user and evaluator').
TA tests concern also non-task issues	Easterners are more likely to have a socio-emotional orientation (Section 'The overall relationship between user and evaluator'). Thus they may perceive the relationship with the evaluator as more than solving tasks or thinking out loud.

Implications for research on TA

The ways in which culture may influence TA testing also raise issues relating to research. In the following discussion we stay within the context of usability evaluation, but the arguments may hold also for the use of TA in general. One implication of our analysis is that it may not be possible to compare or easily transfer the findings of

research based on thinking aloud in one cultural setting to another setting. We, for example, find it problematic to compare local-language interfaces using thinking aloud protocols with local users (say, an English version of a web site tested with English users compared to a Chinese version of the web site tested with Chinese users). It is not clear whether, for example, a difference in strength of preference is due to differences in the success of localization or to cultural differences among test participants. Another example is a study by Hall et al. (2004), who used a procedure for problem discovery based on the test participants' remarks indicating disapproval, surprise, doubt, and so forth. Other studies of usability testing have used similar procedures. Our analysis indicates that the expression of surprise will be higher among Westerners than Easterners. Thus, treating such expressions as indications of usability problems might confound usability issues and cultural differences. Therefore any usability-evaluation research based on counting numbers of usability problems discovered in this way appears difficult to interpret in a cross-cultural perspective.

More research is also needed on what constitutes user and evaluator groups in studies of TA. Nisbett's prototypical samples of test participants (university students from US and Asian universities) are representative in terms of his hypotheses about East-West differences in cognition. In contrast, research on TA often uses students from only one part of the world and without clear criteria for the differences between groups. In their study of cultural dimensions of TA, Hall et al. (2004) used what they saw as two homogenous groups of participants that differed only on Hofstede's cultural dimensions by being either African/Asian or Dutch. Yet, the African/Asian participants were foreign students who at the time of the study were living in Holland, and their answers to a survey could be biased by differences in their inclination to adjust to aspects of the majority Dutch culture. In studies of usability evaluation with inhomogeneous user groups (e.g., Law & Hvannberg, 2004), the borders of the participant groups are typically taken for granted. To be able to assess and appreciate the within-group homogeneity and the between-group heterogeneity we need information about, among other things, the participants' cultural backgrounds.

Our paper has identified a number of areas that need further research so as to characterize more completely the pitfalls and possibilities of cross-cultural TA. First, priming is a strong mechanism for engendering culture-specific cognition. It is not clear, however, whether and how this could be used in TA to dispense with the need for testing with multiple cultures. Second, more research is needed on the effect of thinking aloud on users' task performance during usability tests, and in other contexts where the thinking-aloud method is used. To remove differences in users' ability to think aloud concurrently with performing tasks, Hall et al. (2004) had participants think aloud after the full set of tasks had been completed. Another possibility is to have users think aloud after the completion of each task. It is currently unknown how such variations of TA are influenced by differences in cultural background. Third, the relative benefits of testing with evaluators and users who share or do not share cultural background are largely unknown. Yeo (2001) provides some initial data on this question but other cultures and other combinations of evaluators and users must also be investigated.

Conclusion

The thinking-aloud method is widely used and generally assumed to be applicable irrespective of the cultural backgrounds of the users and evaluators involved in evaluations. It appears, however, that Nisbett's work on cultural cognition challenges this assumption and provides compelling evidence of profound differences in the cognition of Easterners and Westerners. Based on a simplified model of TA as consisting of four main elements, this paper has analysed how culture may affect, in particular, usability evaluations that are based on TA. We find that all four elements of the TA model are susceptible to substantial cultural effects:

- Western users are likely to perceive tasks similarly whether presented as mere instructions or embedded in scenarios. Conversely, Eastern users may find it hard to relate to mere instructions and they attend more to the contextual information provided in scenarios.
- Whereas thinking aloud impairs Easterners' task performance, it appears that Westerners remain relatively unaffected or even improve their task performance. Asking users to provide reasons for their actions draws Westerners toward uncompromising options and Easterners toward middle options.
- Western users are easier for an evaluator to read because they experience and express more surprise and thereby provide more of a record of the problems they experience. Conversely, Eastern users experience less surprise, and it may take a more trained evaluator to detect their problems. Also, Western evaluators

are likely to attribute problems to a few factors, whereas Eastern evaluators emphasize a set of interwoven factors.

- Eastern users and evaluators make many points indirectly and with finesse to maintain social harmony throughout evaluations. Conversely, Westerners attend less to the interpersonal climate of the situation and express themselves more directly, often to the point of condescension. As a result, the overall relationship between user and evaluator becomes substantially more complex when they have different cultural backgrounds.

In research as well as in advice to practitioners, the test participants' cultural backgrounds are frequently neglected as important to the process and results of usability evaluation. This study has argued that to obtain valid results in cross-cultural usability evaluation it is necessary to consider a number of differences in the cognition of people with different cultural backgrounds. While this study has applied a distinction between Easterners and Westerners and found evidence of profound differences between test participants with these two cultural backgrounds, many more differences may exist between test participants with other cultural backgrounds, all of which may eventually be important contributors to the results of a usability test.

Acknowledgements

This study was co-funded by the Danish Council for Independent Research (DCIR) through its support of the Cultural Usability project. For contributions, discussions, comments, and help with earlier drafts of this paper, we owe thanks to Lene Nielsen, Thomas Plocher, Thomas Snitker, and Xianghong Sun.

References

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

Briley, D. A., Morris, M. W., and Simonson, I. "Reasons as carriers of culture: Dynamic versus dispositional models of cultural influence on decision making," *Journal of Consumer Research* (27:2), 2000, pp. 157-178.

Brown, D. E. *Human Universals*, McGraw-Hill, New York, 1991.

Callahan, E. "Interface design and culture," *Annual Review of Information Science and Technology* (39), 2005, pp. 257-310.

Chavan, A. L. "Another culture, another method," in *Proceedings of the 11th International Conference on Human-Computer Interaction*. Erlbaum, Mahwah, NJ, 2005.

Choi, I., and Nisbett, R. E. "Situational salience and cultural differences in the correspondence bias and actor-observer bias," *Personality and Social Psychology Bulletin* (24:9), 1998, pp. 949-960.

Choi, I., and Nisbett, R. E. "Cultural psychology of surprise: Holistic theories and recognition of contradiction," *Journal of Personality and Social Psychology* (79:6), 2000, pp. 890-905.

Chua, H.F., Leu, J., and Nisbett, R.E. "Culture and diverging views of social events," *Personality and Social Psychology Bulletin* (31:7), 2005, pp. 925-934.

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

Cordes, R. E. "Task-selection bias: A case for user-defined tasks," *International Journal of Human-Computer Interaction* (13:4), 2001, pp. 411-419.

del Galdo, E., and Nielsen, J. (Eds.). *International User Interfaces*, Wiley, New York, 1996.

Dumas, J. S. "User-based evaluations, " in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, Jacko, J. A., and Sears, A. (Eds.), Erlbaum, Mahwah, NJ, 2003, pp. 1093-1117.

Dumas, J. S., and Redish, J. C. *A Practical Guide to Usability Testing. Revised Edition*, Intellect, Exeter, UK, 1999.

Ericsson, K. A., and Simon, H. A. "Verbal reports as data," *Psychological Review* (87:3), 1980, pp. 215-251.

Ericsson, K. A., and Simon, H. A. *Protocol Analysis. Verbal Reports as Data. Revised Edition*, MIT Press, Cambridge, MA, 1993.

Evers, V. "Cross-cultural applicability of user evaluation methods: A case study amongst Japanese, North-American, English and Dutch users," in *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, ACM Press, New York, 2002, pp. 740-741.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., and Herulf, L. "Making a difference - a survey of the usability profession in Sweden," in *Proceedings of the Third Nordic Conference on Human-Computer Interaction*, ACM Press, New York, 2004, pp. 207-215.

Hall, M., de Jong, M., and Steedhouder, M. "Cultural differences and usability evaluation: Individualistic and collectivistic participants compared," *Technical Communication* (51:4), 2004, pp. 489-503.

Haritos, C., and Nelson, K. "Bilingual memory: The interaction of language and thought," *Bilingual Research Journal* (25:4), 2001, pp. 417-438.

Hertzum, M., and Jacobsen, N. E. "The evaluator effect: A chilling fact about usability evaluation methods," *International Journal of Human-Computer Interaction* (15:1), 2003, pp. 183-204.

Hofstede, G. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations*, Sage Publications, Thousand Oaks, CA, 2001.

Hong, Y.-y., and Mallorie, L. M. "A dynamic constructivist approach to culture: Lessons learned from personality psychology," *Journal of Research in Personality* (38:1), 2004, pp. 59–67.

ISO 9241. *Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs)-Part 11: Guidance on Usability*, International Organization for Standardization, Genève, CH, 1998.

Jacobsen, N. E., Hertzum, M., and John, B. E. "The evaluator effect in usability studies: Problem detection and severity judgments," in *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, HFES, Santa Monica, CA, 1998, pp. 1336–1340.

Ji, L.-J., Peng, K., and Nisbett, R.E. "Culture, control, and perception of relationships in the environment," *Journal of Personality and Social Psychology* (78:5), 2000, pp. 943-955.

Ji, L.-J., Zhang, Z., and Nisbett, R. E. "Is it culture or is it language? Examination of language effects in cross-cultural research on categorization," *Journal of Personality and Social Psychology* (87:1), 2004, pp. 57-65.

Kim, H. S. "We talk, therefore we think? A cultural analysis of the effect of talking on thinking," *Journal of Personality and Social Psychology* (83:4), 2002, pp. 828-842.

Kühnen, U., Hannover, B., Röder, U., Shah, A.A., Schubert, B., Upmeyer, A., and Zakaria. S. "Cross-cultural variations in identifying embedded figures: Comparisons from the United States, Germany, Russia and Malaysia," *Journal of Cross-Cultural Psychology* (32:3), 2001a, pp. 366-372.

Kühnen, U., Hannover, B., and Schubert, B. "The semantic-procedural interface model of the self: The role of self-knowledge for context-dependent versus context-independent modes of thinking," *Journal of Personality and Social Psychology* (80:3), 2001b, pp. 397-409.

Lavery, D., Cockton, G., and Atkinson, M. P. "Comparison of evaluation methods using structured usability problem reports," *Behaviour & Information Technology* (16:4), 1997, pp. 246-266.

Law, E. L.-C., and Hvannberg, E. T. "Analysis of combinatorial user effect in international usability tests," in *Proceedings of the CHI 2004 Conference on Human Factors in Computing Systems*, ACM Press, New York, 2004, pp. 9-16.

Lewis, C. *Using the 'Thinking-Aloud' Method in Cognitive Interface Design*, Report no. RC 9265 #40713, IBM Thomas J Watson Research Center, Yorktown Heights, NY, 1982.

Lewis, J. R. "Sample sizes for usability studies: Additional considerations," *Human Factors* (36:2), 1994, pp. 368-378.

Lewis, J. R. "Usability testing," in *Handbook of Human Factors and Ergonomics*, Salvendy, G. (Ed.), Wiley, New York, 2006, pp. 1275-1316.

Ling, C. *Advances in Heuristic Evaluation*, Purdue University Dissertation, 2005, http://docs.lib.purdue.edu/dissertations/AAI3210740/.

Masuda, T., and Nisbett, R. E. "Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans," *Journal of Personality and Social Psychology* (81:5), 2001, pp. 922-934.

Miller, J. G. "Culture and the development of everyday social explanation," *Journal of Personality and Social Psychology* (46:5), 1984, pp. 961-978.

Miyamoto, Y., Nisbett, R. E., and Masuda, T. "Culture and the physical environment - holistic versus analytic perceptual affordances," *Psychological Science* (17:2), 2006, pp. 113-119.

Monk, A., Wright, P., Haber, J., and Davenport, L. *Improving Your Human-Computer Interface: A Practical Technique*, Prentice Hall, New York, 1993.

Murphy, J. "Modelling designer-tester-subject relationships in international usability testing," in *Designing for Global Markets. International Workshop on Internationalization of Products and System* (Vol. 3), Day, D., and Dunckley, L. (Eds.), IWIPS, Milton Keynes, UK, 2001, pp. 33-44.

Nielsen, J. *Usability Engineering*, Academic Press, London, 1993.

Nielsen, J., Clemmensen, T., and Yssing, C. "Getting access to what goes on in people's heads? - Reflections on the think-aloud technique," in *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, ACM Press, New York, 2002, pp. 101-110.

Nisbett, R. E. *The Geography of Thought: How Asians and Westerners Think Differently – and Why*, Brealey, London, 2003.

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

Nørgaard, M., and Hornbæk, K. "What do usability evaluators do in practice? An explorative study of think-aloud testing," in *Proceedings of the DIS 2006 Conference on Designing Interactive Systems*, ACM Press, New York, 2006, pp. 209-218.

Page, C., and Rahimi, R. "Concurrent and retrospective verbal protocols in usability testing: Is there value added in collecting both?" in *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, HFES, Santa Monica, CA, 1995, pp. 223-227.

Peng, K. P., and Nisbett, R. E. "Culture, dialectics, and reasoning about contradiction," *American Psychologist* (54:9), 1999, pp. 741-754.

Pinker, S. "The blank slate," *The General Psychologist* (41:1), 2006, pp. 1-8.

Preece, J., Rogers, Y., and Sharp, H. *Interaction Design: Beyond Human-Computer Interaction. Second Edition*, Wiley, New York, 2007.

Rubin, J. (1994). Handbook of usability testing: How to plan, design and conduct effective tests. New York: Wiley.

Sahay, S., Nicholson, B., and Krishna, S. *Global IT Outsourcing: Software Development across Borders*, Cambridge University Press, Cambridge, UK, 2003.

Sanchez-Burks, J., Lee, F., Choi, I., Nisbett, R., Zhao, S., and Koo, J. "Conversing across cultures: East-West communication styles in work and nonwork contexts," *Journal of Personality and Social Psychology* (85:2), 2003, pp. 363-372.

Sanchez-Burks, J., Nisbett, R. E., and Ybarra, O. "Cultural styles, relational schemas, and prejudice against out-groups," *Journal of Personality and Social Psychology* (79:2), 2000, pp. 174-189.

Smagorinsky, P. "Thinking and speech and protocol analysis," *Mind, Culture, and Activity* (5:3), 1998, pp. 157-177.

Smith, A., Dunkley, L., French, T., Minnocha, S., and Chang, Y. "A process model for developing usable cross-cultural websites," *Interacting with Computers* (16:1), 2004, pp. 63-91.

Smith, A., and Yetim, F. (Eds.). "Global human-computer systems: Cultural determinants of usability" (special issue), *Interacting with Computers* (16:1), 2004.

van Baaren, R. B., Horgan, T. G., Chartrand, T. L., and Dijkmans, M. "The forest, the trees, and the chameleon: Context dependence and mimicry," *Journal of Personality and Social Psychology* (86:3), 2004, pp. 453-460.

van den Haak, M. J., de Jong, M. D. T., and Schellens, P. J. "Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison," *Interacting with Computers* (16:6), 2004, pp. 1153-1170.

Vatrapu, R., and Pérez-Quiñones, M. "Culture and usability evaluation: The effects of culture in structured interviews," *Journal of Usability Studies* (1:4), 2006, pp. 156-170.

Vermeeren, A. P. O. S., van Kesteren, I. E. H., and Bekker, M. M. "Managing the evaluator effect in user testing," in *Proceedings of the INTERACT '03 Conference on Human-Computer Interaction*, IOS Press, Amsterdam, 2003, pp. 647-654.

Wang, L. "Switching to first language among writers with differing second-language proficiency," *Journal of Second Language Writing* (12:4), 2003, pp. 347-375.

Wang, W., and Wen, Q. "L1 use in the l2 composing process: An exploratory study of 16 Chinese EFL writers," *Journal of Second Language Writing* (11:3), 2002, pp. 225-246.

Witkin, H. A., Moore, C. A., Goodenough, D. R., and Cox, P. W. "Field-dependent and field-independent cognitive styles and their educational implications," *Review of Educational Research* (47:1), 1977, pp. 1-64.

Yeo, A. W. "Global-software development lifecycle: An exploratory study," in *Proceedings of the CHI 2001 Conference on Human Factors in Computing Systems*, ACM Press, New York, 2001, pp. 104-111.