

SUMMARY

- ◆ Compares the behavior and attitudes of participants from different cultures in two types of usability tests
- ◆ Shows that retrospective think-aloud protocols are less affected by cultural differences than the plus-minus method

Cultural Differences and Usability Evaluation: Individualistic and Collectivistic Participants Compared

MARINDA HALL, MENNO DE JONG, AND MICHAËL STEEHOUDER

INTRODUCTION

As with any form of international communication, Web site designers must take their audience into consideration and respect cultural differences. Consequently, there is a growing research interest in the international and multicultural aspects of Web communication. So far, studies have addressed the relationship between cultural dimensions and the adoption of the Internet (for example, La Ferle, Edwards, and Mizuno 2002), the way cultural background affects people's use and appreciation of Web sites (for example, O'Keefe and colleagues 2000; Simon 2001), and the way both local and international organizations deal with cultural issues on their Web sites (for example, Marcus and Gould 2000; Becker 2002; Okazaki and Rivas 2002). Arnold (1998) gives an overview of the linguistic, cultural, legal, and technical challenges that Web site designers have to face when addressing international audiences. The available research suggests that cultural differences are indeed a relevant factor to consider for Web site designers.

An obvious approach to optimizing a Web site for users from various cultural backgrounds would be to evaluate the site with potential users from all nationalities (see Hoft 1995; Nielsen 2000). Various methods are available for the evaluation of Web sites (see Schriver 1989; de Jong and Schellens 1997), as well as several textbooks with detailed instructions for such tests (such as Rubin 1994; Dumas and Redish 1999; Barnum 2002; Schweibenz and Thissen 2003).

As a result of reviewing methodological research in this area, we can conclude that an evaluation with potential users is an effective way of monitoring and improving the usability of documents and interfaces (de Jong and Schellens 2000). Both *in-use* evaluation methods, such as think-

aloud usability tests, and *non-use* approaches based on verbal self-reports of participants—for example, using the plus-minus method—appear to provide useful feedback for usability engineers and technical writers.

However, a question that has not yet been addressed is to what extent the participants' cultural backgrounds may affect the process and results of a usability study. Is a think-aloud usability test a similar experience for, say, West European and Asian participants, and does it yield comparable results for both groups of participants? And how does each group of participants behave in a plus-minus evaluation study? A large amount of research into the merits and restrictions of evaluation methods has been conducted in North America and Western Europe. Whether the findings of these studies also apply to other cultures is unclear. It is easily imaginable that some methods are better suited for a specific culture than others.

So far, relatively little research has been conducted into the effects that participants' background characteristics have on the feedback collected in document or interface evaluations (see de Jong and Schellens 2000). The available studies have focused on gender and educational level (de Jong and Schellens 2001), and expertise and prior knowledge (Diamantopoulos, Reynolds, and Schlegelmilch 1994). In an analysis of plus-minus evaluation feedback on four brochures, de Jong and Schellens (2001) found that male and female participants provided different feedback, varying by brochure topic. They also found that highly educated participants found more problems in a brochure

Manuscript received 12 March 2003; revised 2 April 2004; accepted 3 April 2004.

than participants with a lower level of education, and focused more on problems in the structure of the information. In an experimental study into the pretesting of questionnaires, Diamantopoulos, Reynolds and Schlegelmilch (1994) found that prior knowledge and expertise in the field of questionnaire design helped participants detect various types of problems in a questionnaire (such as ambiguous questions or missing response alternatives). Apparently, the types of participants recruited may influence the results of a reader-focused evaluation.

Given the international use and possibilities of the World Wide Web, it does seem worthwhile to explore the effects of national culture as a potentially important background characteristic of participants. After all, national culture has proven to be a major influence on people's behavior in many respects (see Hall 1977; Hofstede 1994; Smith and Bond 1998; Trompenaars and Hampden-Turner 1998). This finding leads to the main research question of this study: *Does cultural background influence the feedback collected in and the process of a Web evaluation study?* To answer this question, we conducted a Web usability study among West European and Asian/African participants, and compared the results as well as the participants' experiences.

Of the gamut of possible evaluation approaches, we focused on two current methods: a usability test with retrospective think-aloud protocols, and a plus-minus evaluation. The two methods differ on the *in-use versus non-use* dimension. Retrospective think-aloud protocols are a typical *in-use* approach. Participants carry out tasks with a Web site and their behavior is recorded on videotape; in the second part of the sessions, the participants watch their video recording and try to verbalize the thoughts they had during task execution (see Nielsen 1993). We chose retrospective instead of concurrent think-aloud protocols because of the multilingual sample of participants we recruited; the research was conducted in English, which was not the mother tongue of any of the participants. We assumed that concurrent verbalization would place too high a demand on the cognitive skills of the participants. Furthermore, the few comparative studies available suggest that concurrent and retrospective verbal protocols yield more or less similar results (Hoc and Leplat 1983; van den Haak, de Jong, and Schellens 2003, forthcoming).

The plus-minus method is a typical *non-use* method (see de Jong 1998; de Jong and Schellens 1998). Participants read a document and put pluses and minuses in the margins to anchor positive and negative reading experiences of varying kinds. In the second part of each session, the participants are interviewed about the reasons for the pluses and minuses they recorded. Although some attempts have been made to adapt the method for online Web site evaluation (see Sienot 1997), we asked partici-

pants to evaluate paper copies of some of the Web pages.

The two methods were selected as state-of-the-art representatives of current evaluation techniques. The combination is interesting in the context of intercultural research since both methods require a high degree of participant interaction but place very different demands on a participant. In a think-aloud usability test, participants are expected to *act* as real users, and give insight into the mistakes they make and the doubts they have in the process. With the plus-minus method, participants are expected to *judge* a Web site, and explain their judgments to the facilitator. Both demands may be threatening to some extent to usability test participants.

DIMENSIONS OF CULTURAL DIFFERENCES

Cultural differences may be characterized using so-called cultural dimensions—that is, aspects of cultures that can be measured relative to other cultures. An influential set of cultural dimensions was developed by Hofstede (1994, 2001). Based on a survey of 116,000 IBM employees across 50 countries and 20 languages, he distinguished a number of cultural dimensions.

Masculinity versus femininity

Masculine cultures focus strongly on achievement, assertiveness, and material success, whereas feminine cultures put more emphasis on relationships, caring, and quality of life. Another aspect of this dimension is that male and female roles are clearly distinguished in masculine cultures, while there is less role differentiation between genders in feminine cultures.

Strong versus weak uncertainty avoidance

In cultures with a strong uncertainty avoidance, people feel easily threatened by uncertain or unknown situations. In cultures with a weak uncertainty avoidance, people are willing to try new things, take risks, and accept dissenting views.

High versus low power distance

In cultures with a high power distance, it is accepted and expected that there are differences in power and wealth among people. Cultures with a low power distance, on the other hand, value equal rights and opportunities for everyone.

Individualism versus collectivism

Individualism pertains to societies in which ties are loose: people are expected to look after themselves and their immediate family. Collectivism pertains to societies in which people are integrated into strong, cohesive groups from birth onwards, protecting them in exchange for unquestioned loyalty.

Long-term versus short-term orientation

This dimension reflects the extent to which a society has a pragmatic and future-oriented perspective rather than a conventional, historic, or short-term point of view. A culture with a high long-term orientation values perseverance and thrift, prioritizes general purposes over individual interests, and orders relationships by status. A culture with a short-term orientation, on the other hand, focuses on quick results, and honors tradition, personal standpoints, social obligations, and people's need to protect their "face."

Cross-cultural studies have provided empirical evidence that the individualism-collectivism dimension is the most important dimension for pinpointing differences between cultures (see Ting-Toomey 1998). Various authors claim that this dimension is particularly related to the way people communicate with each other, which is reflected by Hall's (1977) distinction between low-context and high-context cultures (Hofstede 2001; Ting-Toomey 1998). In low-context cultures, communication is expected to be explicit, direct, and unambiguous. In high-context cultures, on the other hand, most information is either part of the context or is internalized in the persons involved; very little is made explicit as part of the message. High-context communication corresponds to the collectivist culture, while low-context communication fits the individualist society. "Many things that are self-evident must be said explicitly in individualist cultures" (Hofstede 2001, p. 212).

In our view, the cultural differences between West European and Asian/African participants can best be characterized by a combination of Hofstede's individualism-collectivism dimension and Hall's distinction between low-context and high-context cultures. According to the available cultural indexes, our Asian/African participants might be expected to be on the collectivism/high-context end of the continuum, and our West European participants on the individualism/low-context end. It must be stressed, however, that the dimensions reflect group differences, and cannot be used to predict individual behaviors. It is also important to realize that cultures often differ in more than one dimension at the same time. So our classification of the two groups of participants is an intentional simplification of reality, which, as we will show further on, proved to be a fruitful basis to develop research hypotheses for this particular study.

CULTURE AND POLITENESS THEORY: RESEARCH HYPOTHESES

Ting-Toomey (1998) connected theories about cultural differences to Brown and Levinson's politeness theory (1990). A central notion in this theory is people's desire to maintain their face, "the public self-image that every member of a society wants to claim for himself" (p. 61). People want to be appreciated by others (positive face) and do not want to be forced by others to do things they do not want to do

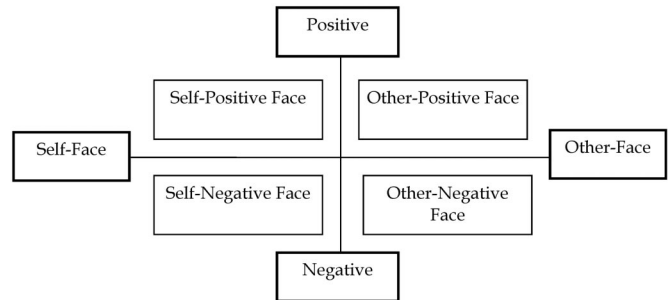


Figure 1. Two-dimensional grid of facework maintenance (Ting-Toomey 1998, p. 218).

(negative face). Brown and Levinson have developed a typology of possible face-threatening acts, and have explored the various ways in which they can be performed.

Examples of positive-face-threatening acts include expressions of disapproval and criticism—the speaker (S) indicates that he/she does not like or want one or more of the hearer's (H) needs, goods, or personal characteristics. Negative-face-threatening acts include orders—S indicates that he/she wants H to do something. And even offers may be negative face-threatening acts—S commits him/herself to a future act for H's benefit, putting pressure on H to accept or reject the offer and possibly to incur a debt. Although the interpretation of face may differ between cultures, the importance of the concept in human interaction is universally acknowledged.

Based on the distinction between self-face concern and other-face concern, and the above-mentioned distinction between positive and negative face, Ting-Toomey (1998) developed a theoretical framework. Figure 1 shows how these concepts are used to form a two-dimensional grid. People trying to maintain self-positive face use communication strategies to defend and protect their need for inclusion and appreciation. Other-positive face maintenance includes strategies to maintain, defend, and support another persons' need for inclusion and appreciation. People trying to maintain self-negative face use interaction strategies to give themselves freedom and space, and to protect themselves from infringements on their autonomy. Other-negative face maintenance indicates the use of interaction strategies to show respect for other persons' need for freedom, space, and disassociation.

According to Ting-Toomey, individualistic/low-context cultures differ from collectivistic/high-context cultures in many respects. Using the framework in Figure 1, she developed a set of theoretical propositions indicating the major, facework-related characteristics of the two types of cultures. These propositions form the basis of the re-

TABLE 1: INDIVIDUALISTIC/LOW-CONTEXT VS. COLLECTIVISTIC/HIGH-CONTEXT FACEWORK (BASED ON TING-TOOMEY (1998, p. 230))

Key elements of “face”	Individualistic/low-context	Collectivistic/high-context
Identity	Emphasis on “I” identity	Emphasis on “we” identity
Concern	Self-face concern	Other-face concern
Need	Negative face need	Positive face need
Supra-strategy	Self-positive and self-negative facework	Other-positive and other-negative facework
Mode	Direct mode	Indirect mode
Style	Controlling, confrontational, solution-oriented style	Obliging, avoiding, and affective-oriented style
Speech acts	Direct speech acts	Indirect speech acts
Nonverbal acts	Individualistic nonverbal acts, direct emotional expressions	Contextualistic (role-oriented) nonverbal acts, indirect emotional expressions

search hypotheses of this study. The propositions are summarized in Table 1.

The first question addressed in this article is whether the participants’ cultural background affects the *results* of a usability evaluation. Based on Ting-Toomey’s assumptions about differences in facework between the two types of cultures, it seems plausible that the results of a plus-minus evaluation might be affected by the participants’ cultural background. Participants in a plus-minus evaluation are supposed to provide direct feedback on the Web site, a process that does not match with the indirect mode of communication that is dominant in collectivistic/high-context cultures, and may not be compatible with people’s other-face concern. For the retrospective think-aloud protocols, effects of cultural differences between participants are less likely. The majority of the problems found with this method are related to mistakes and doubts that occur during the process of using the Web site. There is no reason beforehand to assume that one of the two groups of participants will experience more usability problems during a session. After all, the two groups of participants recruited in our study have the same educational and work level (see the description of participants later in this article). As a result, we formulated the following two research hypotheses regarding the effects of cultural background on evaluation results.

H1. The plus-minus method will reveal fewer usability problems when used by participants from collectivistic/high-context cultures than by participants from individualistic/low-context cultures.

H2. The retrospective think-aloud protocols will reveal an equal number of usability problems when used by participants from collectivistic/high-context and by participants from individualistic/low-context cultures.

The participants’ cultural background may also affect the *experiences* of participants during the evaluation session. There may be a match between the kind of facework preferred in the two cultures and the way participants experience a particular evaluation method. In collectivistic/high-context cultures, people have a dominant other-face concern, and a positive face need. Both may be threatened by the requirement for participants to openly criticize a Web site. It may therefore be assumed that the plus-minus method would be less appreciated by participants from collectivistic/high-context cultures than by participants from individualistic/low-context cultures. In individualistic/low-context cultures, on the other hand, people have a dominant self-face concern, and a negative face need. Think-aloud protocols might threaten the participants’ self-face because they expose all things that go wrong in the process of using the Web site (although the participants might, of course, assign the blame for the problems they encounter to the Web site instead of to

themselves; see Hypothesis 5). The negative face need may also be problematic, since participants are forced to use the Web site in certain preset ways (although the participants may decide to take a broader definition of their participant role; see Hypothesis 6). As a result, it may be assumed that the retrospective think-aloud method will be less appreciated by participants from individualistic/low-context cultures than by participants from collectivistic/high-context cultures. These assumptions lead to two more hypotheses regarding participant experiences.

H3. Participants from collectivistic/high-context cultures will judge the plus-minus method less favorably than participants from individualistic/low-context cultures.

H4. Participants from individualistic/low-context cultures will judge retrospective think-aloud protocols less favorably than participants from collectivistic/high-context cultures.

A third aspect that may be of interest from a cultural perspective concerns the *attribution of blame*. Schriver (1997) described a study in which people were asked to assign the blame for the usability problems they had when working with consumer products. A majority of her participants assigned the blame to themselves instead of to the device or the manual. The question as to whether participants blame themselves (internal blaming) or the Web site or test situation (external blaming) might be culture-specific. Because of a dominant other-face concern, collectivistic/high-context participants might be inclined to cast the blame for problems on themselves; due to their self-face concern, individualistic/low-context participants might want to assign the blame to external factors.

H5. Participants from collectivistic/high-context cultures will be more inclined to blame themselves for the problems they experience in the retrospective think-aloud test than participants from individualistic/low-context cultures.

A fourth relevant aspect concerns the *role* of the participants during the retrospective think-aloud test. Participants from individualistic/low-context cultures are assumed to have a stronger negative face need than participants from collectivistic/high-context cultures. A retrospective think-aloud test may be too limiting for them: they are supposed to act as users with a specific set of tasks, and primarily report on the task-related problems they encounter. In reaction to this limitation, they might take on other roles than the strict user role that is assumed by the method. Examples of such roles would be the role of test participant (reporting on their experiences in the test situation and commenting on their performance in the test), real-world Internet user (reporting on their normal use of the Web site or the Internet in general), or reviewer (judging the Web site instead of using it).

H6. During the retrospective think-aloud test, participants from collectivistic/high-context cultures will be more inclined to keep to the user role assigned to them than will participants from individualistic/low-context cultures.

The last hypothesis concerns the way participants *express* their criticism during the evaluation session. Participants from collectivistic/high-context cultures would be likely to prefer indirect speech acts when commenting on the Web site because of their other-face concern and their positive face need. Participants from individualistic/low-context cultures would be assumed to be more direct in their comments.

H7. Participants from collectivistic/high-context cultures will be more inclined to use indirect and euphemistic formulations of the problems they encounter than participants from individualistic/low-context cultures.

Note that this hypothesis was tested only in the retrospective think-aloud condition. The plus-minus results contained so many comments and so many combinations of direct and indirect utterances that it was not possible to reliably categorize them.

MATERIALS AND METHODS

To test the seven hypotheses described above, we conducted a Web usability study with two groups of participants (Asian/African and West European PhD students), each of whom was asked to evaluate the Web site using two methods (retrospective think-aloud protocols and the plus-minus method). In this section of the article we describe the Web site selected, the participants, and the procedure we used for the tests, as well as the dependent variables that we tested.

Web site used for evaluation: *Web of science*

For the purpose of our research, we selected a Web site that met the following criteria:

- ◆ **A clear instructive function** For the retrospective think-aloud protocols, participants needed to be able to complete a set of realistic tasks using the Web site.
- ◆ **Substantial textual content** For the plus-minus evaluation, the Web site must contain textual information that participants could read and comment on.
- ◆ **No culture-specific intentions or content** The Web site should focus on user groups from various cultures. Specifically, it had to be equally plausible that both Asian/African and West European participants would visit and use the site.

We decided to use the *Web of science*, a database published and maintained by the Institute for Scientific Information (ISI), for our study. The database enables sci-

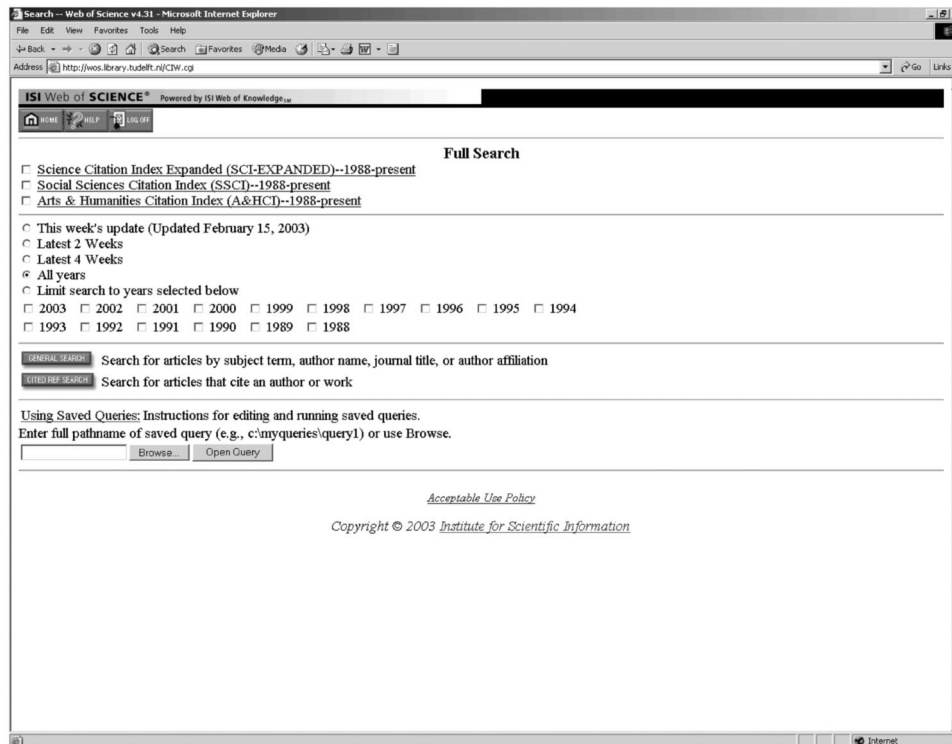


Figure 2. The *Web of science* “full-search” page.

entific researchers and evaluators to look up citation indexes of specific articles. The *Web of science* covers a large number of high-quality scientific journals (*Technical communication* is included) in the domains of science, social sciences, and arts and humanities. It is accessible on a subscription basis. For PhD students, the *Web of science* is an important source of information: it is one of the ways to systematically uncover scientific literature. Figure 2 shows the “full search” screen of the database.

Participants

To focus on cultural differences, we recruited two homogeneous groups of participants, who only differed on the cultural dimension. We asked male PhD students of technical sciences at the University of Twente to participate in our study. This sample appears to be suitable for our research purposes in many respects:

Most of the participants had some experience using the *Web of science* database.

The participants had similar educational backgrounds. All participants had received their master's degree and they were all working for their PhD degree. This fact is important, because prior research has shown that there may be a relationship between educational level and reader feed-

back produced (see de Jong and Schellens 2001).

The collectivistic/high-context participants had all come to the Netherlands only one or two years previously to conduct a PhD research project. They had not been raised in the Netherlands and therefore could be considered to be representatives of their native cultures.

The participants' linguistic skills in English (the language of the *Web of science* site and the language that was used throughout the research sessions) were the same. English was a second language for all participants. Considering the equal level of prior education and the increasing use of English at Dutch universities, it can be assumed that there were no differences in language command between the two groups of participants. This is important because the participants' linguistic skills could influence their verbalizations in the retrospective think-aloud protocols and the number and types of comments given under the plus-minus method.

These selection criteria make it possible to explore the cultural differences in-depth, but, of course, there is also a down side in terms of generalizability: the research findings are restricted to highly-educated male participants.

Participants who met the above-mentioned criteria were asked face-to-face whether they were willing to participate in

1. How many articles are available about the topic "communication theory"? Make sure that you find articles that use these two words simultaneously and consecutively.
2. Look up how many articles are available about "Web evaluation". This time, the two terms do not have to be consecutive.
3. Please save the last search results on the A-drive of this computer.
4. How many articles written by Jan H. Spyridakis are available in the database?
5. Does the database include a journal called *IEEE Transactions on Knowledge and Data Engineering*?
6. How many journals are in the Science Citation Index Expanded?
7. How often has the article by P.J. Schellens in *Technical Communication* been cited in other articles?

Figure 3. Tasks given to the participants.

the study. In total, 38 PhD students agreed to participate: 20 participants were recruited from individualistic/low-context cultures (the Netherlands), and 18 from collectivistic/high-context cultures (India, Indonesia, China, Turkey, and Sudan).

Procedure

The research was conducted in individual sessions in a research facility at the University of Twente. The test consisted of four parts. First, participants were asked to perform seven tasks using the *Web of science*. Their process of using the Web site was recorded using HyperCam (<http://www.hyperionics.com>). Since the kinds of tasks given to participants may influence the findings of a usability test (see Sienot 1997; van Waes 2000), the participants were given seven different tasks, divided into two broad categories: searching and application. To eliminate prior knowledge effects, we chose to formulate tasks that did not correspond to the participants' own research interests (see Figure 3).

After performing the tasks, the participants were asked to view the recordings and think aloud in English about how and why they performed the tasks, stopping the screen recording if necessary. Their verbal accounts were recorded using a tape recorder. Van Someren, Barnard, and Sandberg (1994) argue that data obtained by retrospection is not always valid, especially if there is a delay in time between performing the tasks and explaining them retrospectively. To avoid strong distortions in the data, we asked the participants to view and comment on the recording immediately after they completed the set of tasks.

Next, the participants were placed in an evaluator role and requested to assign pluses and minuses to printouts of a number of help pages from the *Web of science*. These pages explained the purpose and coverage of the *Web of science* and the search possibilities within the site. Participants were told that they could put pluses and minuses in the margin for all conceivable reasons and that they could assign the pluses and minuses to elements of their own choice (varying from individual words to entire pages). When they finished reading the texts and putting pluses

and minuses in the margin, they were asked to explain each plus and minus in an individual interview, which was also tape-recorded.

Finally, the participants completed a questionnaire designed to investigate (1) their experiences during the test; (2) the extent to which they attributed the blame for usability problems to themselves, the Web site or the test situation; and (3) their place on the individualistic-collectivistic continuum. For lack of a better alternative, the last question was addressed using Hofstede's four-item Individualism (IDV) index, although the validity of this index to measure individual cultural differences is questionable (see Hofstede 2001, p. 497).

Dependent variables

To investigate the usability problems detected (H1 and H2), the mean number of problems detected using the plus-minus method and the retrospective think-aloud sessions were computed per participant. For the plus-minus method, all negative remarks made by participants were coded as a problem. For the retrospective think-aloud sessions, problems were indicated either by deviations from the optimal problem-solving process for each task, or by participants' remarks indicating disapproval, surprise, doubt, and so forth.

The participants' experiences with the two methods (H3 and H4) were investigated by three sets of questions using five-point Likert scales. A first set of five questions focused on the participants' experiences with the retrospective think-aloud test (for example, "I felt uncomfortable while performing the tasks"). A second set of four questions focused on the participants' appreciation of the plus-minus method (for instance, "I did not like evaluating the Web site with this method"). And a third set of four questions compared the two evaluation methods (for example, "Evaluating the *Web of science* from paper enabled me to give better feedback for improvements for the Web site than performing the tasks").

The extent to which the participants blamed themselves for the usability problems they encountered while performing the tasks (H5) was investigated by a set of nine questions, again using five-point Likert scales. The participants were asked to judge various explanations for the problems they had. Three of the explanations represented internal blaming (for instance, "The problems I experienced using the database were caused by . . . me not reading correctly"). The other six explanations represented two types of external blaming—that is, addressing the quality of the *Web of science* (for example, ". . . the lack of user-friendliness of the Web site") and the artificial test situation (for instance, ". . . the researcher looking over my shoulder").

To investigate the extent to which the participants kept

to the user role assigned to them in the retrospective think-aloud sessions (H6), we marked all deviations from the default user role in the comments that were made in the think-aloud protocols, distinguishing between the roles of test participant, real-world Internet user, and reviewer (as explained above). In our analysis, we focused on these three specific roles as well as on the total number of non-user roles assumed.

Finally, to investigate the directness in the way participants formulated the problems they found (H7), we made a list of all verbalized problems in the retrospective think-aloud protocols and asked a sample of 12 third-year communication students to rate the directness of each comment on a five-point scale. Together, the twelve students formed a reliable scale (Cronbach's alpha = .82) to distinguish direct and indirect formulations. (Cronbach's alpha determines the reliability of a rating that summarizes a group of test or survey answers used to measure some underlying factor—in this case, the directness of comments.) The analysis was limited to the participants who produced verbal comments on the Web site. The directness scores of the comments were compared using the participants' cultural background as independent variable.

RESULTS

Before addressing the effects of cultural differences on the performance and experiences of the participants, we will briefly sketch some of the results of the usability test itself. Both the think-aloud protocols and the plus-minus method revealed many serious problems in the *Web of science* site. On average, for instance, 33% of the tasks were not completed correctly, with failure percentages varying between 3% for the third assignment (saving search results on a diskette) and 54% for the second assignment (performing a search). The query formats that were required for topic and author searches also caused many problems.

In addition, two peculiarities of the interface appeared to be very contra-intuitive to many of the participants.

1. After participants used the Back button of the browser, the page had to be reloaded.
2. If the user pressed the Enter key after specifying search terms, all search terms were deleted, and the user was directed back to the site's homepage.

These two recurring user problems have by now been repaired on the Web site, but there were also many more specific usability problems. All in all, the *Web of science* appeared to be complicated enough to generate rich think-aloud protocols and plus-minus comments.

One of the collectivistic/high-context participants had to be excluded from our analyses because he did not comply with the requirements of the two evaluation methods. Despite receiving exactly the same instructions as the

other participants, he did not try to carry out the seven tasks in the think-aloud session, and he did not comment on specific aspects of the help page in the plus-minus session. Instead, he explored the *Web of science* and provided an account of how excellent he felt the Web site and the help file were ("It's an exhaustive list and it's definitely very handy to the layman who does not know how and what kinds of words could be entered for a particular search. Indeed very good").

In this section of the article, we will focus our discussion of the results on the seven hypotheses we tested. Before doing so, however, we will address two relevant background characteristics of the participants in the sample—that is, their scores on Hofstede's IDV index, and their prior experiences with the *Web of science*.

Background characteristics of the participants

A first question regarding the participants' background characteristics is whether there is independent support for cultural differences on the individualism-collectivism dimension between the two groups of participants, in addition to their countries of origin. This dimension was measured at the end of each session using the four questions of Hofstede's IDV index. The IDV index did not differ between the two participant groups (*t* test, $t = .554$, $df = 35$, $p = .583$). (A *t* test indicates the probability that the difference between the two means is the result of chance; *t* is the statistic that the test produces; *df* indicates the degrees of freedom; and *p* is the probability. A *p* value of less than .05 is usually considered an indicator that the results are not caused by chance.) The Dutch participants had an IDV score that exactly matched Hofstede's predictions for the Netherlands, but the IDV scores of the Asian and African participants pointed to a more individualistic orientation than was expected based on their national background.

A possible explanation would be the kind of collectivistic/high-context participants recruited for our study. Either by selection or by assimilation, they may have had more individualistic characteristics than we had expected beforehand. Selection bias is possible, since the participants had all made the drastic decision to temporarily move to another country, far away from their home country and different from that country in many respects. Only more or less adventurous people do that, a typical individualistic trait. Assimilation processes are also possible; since all participants had lived in the Netherlands for up to two years, they may have adopted some Dutch views and habits.

Another—in our view, even more important—explanation lies in the fact that the IDV index is not an undisputed measure of cultural differences. The IDV index consists of four items that, surprisingly enough, do not intrinsically correspond to individualistic or collectivistic

TABLE 2: MEAN NUMBER OF PROBLEMS DETECTED PER PARTICIPANT IN THE WEB OF SCIENCE

	Individualistic/ low-context	Collectivistic/ high-context	Significance
Reader problems under the plus-minus method	4.8	1.8	F(1,33) = 8.97, $p < .01$, $\eta^2 = .21$
User problems in the retrospective think-aloud session	7.2	9.5	n.s.

traits and hence, can at best be seen as a cultural predictor. Hofstede (2001, p. 497) himself is not sure of the reliability and validity of the index, in particular when it is used to investigate individual cultural differences. And more generally, there is considerable doubt whether it is possible to bring to light cultural differences by means of written questionnaires (see Peng, Nisbett and Wong 1997).

Based on these considerations, we cannot be absolutely sure that our Asian/African participants fully represent the dominant culture in their native countries. We do know, however, that the two groups differed regarding their origin, and were similar on other possibly relevant characteristics. Since our hypotheses were derived from theories about cultural differences on the individualism-collectivism dimension, it seems reasonable to attribute differences found to the participants' cultural background.

A second background issue concerned the participants' prior experience with the *Web of science*: To what extent were the two groups of participants comparable in this respect? Most of the participants had prior experience with the *Web of science* database. However, there appeared to be an almost significant difference between the individualistic/low-context and collectivistic/high-context participants in this respect: fewer collectivistic participants had used the *Web of science* before the study (59% versus 90%, Fischer's exact test, $p = .052$). (Fischer's exact test is used to determine the independence of variables with small sample sizes.)

The intensity of *Web of science* use, measured by the number of times the participant had used the Web site in the last three months, did not differ between the two groups of participants (1.8 versus 1.9 times, t test, $t = .103$, $df = 26$, $p = .918$). Since prior experience may affect the feedback that participants give (see Diamantopoulos, Reynolds, and Schlegelmilch 1994), we decided to include prior experience (yes/no) in our analyses of the problems detected.

Number of problems detected

Our first two hypotheses dealt with the number of problems detected using the two evaluation methods. We ex-

pected that the collectivistic/high-context participants would mention fewer problems than the individualistic/low-context participants in the plus-minus sessions (H1), and that the number of problems detected by the two groups of participants would not differ in the retrospective think-aloud sessions (H2). As can be seen in Table 2, both hypotheses were confirmed by the results. The η^2 regarding the plus-minus evaluation indicates a substantial difference between the two groups of participants. (The η^2 indicates the percentage of the variance explained by the difference between the two participant groups. An η^2 of .14 or higher is usually considered to indicate a large effect.) Surprisingly, the participants' prior experience with the *Web of science* did not affect the number of problems detected under the two methods, and there was also no interaction effect between cultural background and prior experience.

Besides the detection of specific usability problems, the retrospective think-aloud method also generated an overall success score for each participant on the seven tasks. An analysis of the number of correctly completed tasks did not result in significant differences, neither for the participants' cultural background nor for their prior experience with the *Web of science*.

Appreciation of the two evaluation methods

The next two hypotheses dealt with the participants' appreciation of the two methods. We expected the collectivistic/high-context participants to judge the retrospective think-aloud method more favorably than the individualistic/low-context participants (H3), and the individualistic/low-context participants to be more positive about the plus-minus method (H4). Neither of the two hypotheses was confirmed by our data. The five questions focusing on the participants' appreciation of the retrospective think-aloud method did not form a reliable scale, and had to be analyzed separately. Not one significant difference between the two groups of participants was found in our analysis. The same applies to the four questions about

TABLE 3: ATTRIBUTION OF BLAME FOR USABILITY PROBLEMS ENCOUNTERED (SCORES ON A FIVE-POINT SCALE, 1 = AGREE TO 5 = DISAGREE)

	Individualistic/ low-context	Collectivistic/ high-context	Significance
External blaming: Artificial test situation (mean of three items)	4.4	4.4	n.s.
External blaming: Quality of the Web site (mean of three items)	3.2	3.3	n.s.
Internal blaming: Inexperience with databases	4.4	4.5	n.s.
Internal blaming: Inexperience with Web of science	2.7	2.5	n.s.
Internal blaming: Not reading correctly	3.0	3.3	n.s.

the participants' appreciation of the plus-minus method; they did not form a reliable scale, and there were no differences between the two participant groups. Overall, participants appeared to judge both methods rather favorably.

The four questions comparing the plus-minus method and the retrospective think-aloud method formed two scales with satisfactory Cronbach's alphas (.60 and .67, respectively), focusing on the participants' experiences during the test (easier and more pleasant) and on the perceived usefulness of the feedback provided (more useful information and better feedback for improving the Web site). The results for the two scales were again not significant, but regarding the perceived usefulness, there was a tendency in the opposite direction than expected: collectivistic/high-context participants seemed to judge the plus-minus method more favorably than individualistic/low-context participants (t test, $t = 1.975$, $df = 35$, $p = .056$). There appears to be a discrepancy between the small number of problems detected by the collectivistic/high-context participants and their opinion about the usefulness of the feedback they gave with this method.

Internal or external blame for usability problems

With regard to the roles participants assumed in the retrospective think-aloud test, we expected that the collectivistic/high-context participants would be more inclined to assign blame to themselves for the problems they encountered (H5). We investigated this assumption using nine questions in the questionnaire at the end of the ses-

sion, three of which focused on internal blaming and six on external blaming (that is, blaming the quality of the Web site and the test situation).

The questions about external blaming formed two sufficiently reliable scales (Cronbach's alpha = .64 for Web site quality and .61 for the artificial test situation); the questions about internal blaming did not. As can be seen in Table 3, there were no significant differences regarding blame attribution between collectivistic/high-context and individualistic/low-context participants. Unlike Schriver's (1997) earlier findings, which suggest that usability test participants are very critical about their own performance, the participants in our study did not predominantly blame themselves for the problems they encountered. The participants' estimation that the artificial test situation was not an important cause for the errors they made supports the design of our retrospective think-aloud study: according to the participants, the given tasks were realistic, and the test situation was unobtrusive.

Roles during the retrospective think-aloud method

Another hypothesis focused on the participants' behavior during the retrospective think-aloud test. We expected that the collectivistic/high-context participants would permit themselves fewer role changes than the individualistic/low-context participants (H6). We distinguished among three non-user roles that participants could assume: test participant, real-world Internet user, and reviewer. Examples of utterances from our dataset can be found in Figure 4.

<p>Utterances from a test participant role:</p> <p>I was doing a few checks. That's why I didn't complete all the questions. I learned something from this today. I should pay more attention. It's a stupid way of doing it. But it's the only way I know. I'm getting annoyed with myself. Don't look at this. This is so stupid. I even typed it wrong!</p> <p>Utterances from a real-world Internet user role:</p> <p>Normally, I never use save. Normally, I would ask someone else whether they know how to do it. I always press back on the browser, not on the page. That's what I usually do. In my field of research ...</p> <p>Utterances from a reviewer role:</p> <p>I think this shouldn't be so strict. That's not very good that you have to use a list. How difficult can it be to implement... That was the main problem: enter. It's not one of my favorite databases.</p>

Figure 4. Examples of utterances representing different roles.

Table 4 presents the results of our analysis. In accordance with our hypothesis, participants from individualistic/low-context cultures appeared to be more inclined to assume non-user roles during the retrospective think-aloud test than participants from collectivistic/high-context cultures (with Cohen's *d* indicating a medium to large effect). (Cohen's *d* is a measure to determine the effect size of the difference between the two groups of participants. A Cohen's *d* of .50 is generally considered to indicate a medium effect, a Cohen's *d* of .80 refers to a large effect). This finding, however, applies only to the total number of non-user roles assumed; when the three specific non-user roles were analyzed separately, no significant differences were found. During a think-aloud usability test, individualistic/low-context participants were likely to produce a broader range of problems and observations than collectivistic/high-context participants, while the collectivistic/high-context participants tended to keep more to the user roles that were implied by the tasks they were given.

Direct and indirect comments during evaluation

The final hypothesis that we tested concerned the way participants described the problems they detected: we expected participants from collectivistic/high-context cultures to use more indirect and euphemistic problem descriptions in the retrospective think-aloud test than participants from individualistic/low-context cultures (H7).

A first remarkable finding concerns the number of comments made by the two participants groups. In line with our hypothesis, the vast majority of the comments were given by participants from individualistic/low-context cultures (69 *vs.* 30 percent). This finding corresponds to the hypothesis of role changes described above.

Figure 5 shows some examples of direct and indirect

comments, together with the mean directness scores for each sample comment given by the 12 Communication Studies students (1 = very direct; 5 = very indirect). There appeared to be a difference between the two groups of participants in the directness of the responses, confirming our hypothesis. The comments made by participants from collectivistic/high-context cultures had a mean directness score of 2.8; the comments made by participants from individualistic/low-context cultures resulted in a mean score of 2.2. The difference is not only statistically significant (*t* test, $t = 2.507$, $df = 37$, $p < .05$) but also indicates a large effect (Cohen's $d = .89$).

DISCUSSION

In this section we will first draw conclusions about the effects that cultural differences in the collectivistic/high-context versus individualistic/low-context dimension have on the feedback collected in a Web usability evaluation. Then we discuss the remarkable difference in results between the participants' actual behavior and their verbal self-reports in questionnaires. Finally, we will reflect on our experiences, and address the difficulties of conducting intercultural research.

Cultural influences on usability test results

The overall conclusion from our research is that cultural background of the participants is indeed a relevant variable that may influence the feedback collected in usability evaluation research. Although the IDV index results did not confirm the cultural differences between the two groups of participants in our study, the participants' behavior to a great extent corresponded to our expectations, which were based on the literature about cultural differences. Advice to conduct user research to monitor and improve international Web sites, documents, or interfaces is still valid and useful, but it must be amended by the observation that the evaluation *methods* used may also be susceptible to cultural bias.

The plus-minus method, in particular, seems to be far more useful in individualistic/low-context cultures than in collectivistic/high-context cultures. After all, the Asian/African participants produced dramatically less feedback on the *Web of science* pages than the West European participants. Remarkably enough, this was not reflected in the participants' own judgments about the usefulness of the plus-minus evaluation: the collectivistic/high-context participants tended to be rather positive about the usefulness of their plus-minus feedback. In the available research so far, all conducted in Europe and North America, the plus-minus method has proven to be a very useful method producing abundant feedback on documents (see de Jong 1998). Based on the results we obtained in this study, it seems unlikely that the same conclusion can be drawn when the method is used in a collectivistic/high-context culture.

TABLE 4: NON-USER ROLES ASSUMED BY THE PARTICIPANTS IN THEIR THINK-ALOUD COMMENTS

	Individualistic/ low-context	Collectivistic/ high-context	Significance
Test participant	3.1	2.2	n.s.
Real-world Internet user	2.1	1.1	n.s.
Reviewer	1.4	0.7	n.s.
Total of non-user roles	6.6	4.0	t-test (one-sided), $t = 1.818$, $df = 35$, $p < .05$, Cohen's $d = .61$

Direct responses:

- Stupid thing (1.1).
- That's very annoying (1.2).
- Doesn't work properly. Annoying (1.3).
- Of course, it still didn't work (1.4)

Indirect responses:

- There's something strange ... I guess it's not important (3.5)
- It's not the most convenient way ... (3.3)
- I think this shouldn't be so strict (3.2)
- It's not one of my favorite databases (3.2)

Figure 5. Examples of direct and indirect responses.

Retrospective think-aloud protocols appear to be less susceptible to cultural influences, since the mistakes participants make while working with the Web site are the method's backbone and there are no reasons beforehand to assume that participants from one culture make more mistakes than participants from another culture. Still, we also found two effects of cultural background on the problems participants mentioned.

First, participants from collectivistic/high-context cultures appeared to be more inclined to keep to the user role in their comments on the *Web of science*. While the participants from individualistic/low-context cultures frequently decided to provide comments from other perspectives—criticize the Web site as a reviewer, comment on the test situation, or reflect on things they would do with similar applications in real life—the collectivistic/high-context participants tended to adopt the user role assigned to them and behave accordingly. One could say that they provided “more genuine” think-aloud protocols than the individualistic/low-context participants. On the other hand, the non-user remarks made by the individualistic/low-context participants may also be very valuable to the Web designer.

Second, the participants from collectivistic/high-context cultures formulated their comments less directly than the participants from individualistic/low-context cultures. Such variations in the way problems are expressed might unfairly affect the severity estimation that is attached to the detected problems in the revision phase.

On a more theoretical level, the distinction between collectivistic/high-context and individualistic/low-context cultures, as operationalized by Ting-Toomey (1998), proves to be a fruitful approach to characterize the differences between West European and Asian/African participants. Not all of our hypotheses were confirmed by our data, but for four hypotheses (three referring to a difference between the two groups of participants, and one to similar results), we found substantial empirical support.

Differences between behavior and self-reports

A remarkable discrepancy in our data concerns the difference between behavioral data and verbal self-reports. The three hypotheses that were not confirmed in our study (that is, the hypotheses regarding the participants' experiences with the two evaluation methods, and blame attribution) were investigated using self-report questions. The three differences that were confirmed between the two participant groups (that is, the number of problems reported under the plus-minus method, the participants' tendency to keep to the user role assigned to them, and the directness of the feedback) were all based on behavioral data.

This discrepancy may be explained by Hofstede's (1994) “onion diagram” of culture, which states that people have several layers of culture—with values as the core of a culture, and symbolic behavior as the outer layer. One could argue that the participants' behavior during a usability test reflects a deeper level of the cultural system than the answers they give in a questionnaire. These answers may have been biased by social desirability, which in this study

may reflect the participants' inclination to adjust to aspects of the Dutch culture. Social desirability, in turn, may be connected with cultural differences as well: Middleton and Jones (2000) showed that social desirability is of greater influence in collectivistic cultures than in individualistic cultures.

With hindsight, the influence of social desirability may also serve to explain the non-discriminating IDV scores for the two groups of participants. After all, Hofstede's IDV index is also based on participants' self-reports. The consistent lack of significant cultural differences in all comparisons based on self-report data, as opposed to those based on behavioral data, cast serious doubt on the importance of the IDV scores in this study.

Implications for intercultural communication research

Our experiences question two common practices in intercultural research. Due to possible selection and assimilation effects, recruiting immigrated participants to represent their homeland culture may be questionable, at least when the research is focusing on world-wide cultural differences instead of cultural differences within a national context. An example of the latter would be Lentz and Hulst's (2000) study into the appreciation and use of Dutch public information brochures among, for instance, Moroccan and Turkish immigrants.

A second common practice that is, in our view, questionable is the use of questionnaires, such as the IDV index, to establish the cultural differences between groups of participants: such instruments may focus only on the outer layers of culture and neglect the core values and habits that affect participants' behavior.

Apart from confirming the relevance of cultural differences for modern communication research, the research reported in this article has also made us aware of the complexity of cross-cultural research. Cultural differences may be of interest at many different levels. The well-known Russian matruschka dolls may be an instructive comparison: opening the first doll will reveal a

second one, identical but somewhat smaller; opening the second doll will reveal a third one, and so forth. Our study was an attempt to open the second matruschka doll, but working on it, we became aware of the third and the fourth.

As we mentioned earlier, the research on cultural differences in social desirability may be a factor of interest in interpreting our findings. Another study into the cultural effects on international usability testing focused on the effects of cultural congruency between participants and facilitator: if the facilitator shared the participants' cultural background, the usability interviews appeared to be more productive (see Vatrapu 2002). In our study, we did not include this factor; the collectivistic/high-context participants were interviewed by a Dutch facilitator. Again, this may have influenced the results of our study. Including this congruency in a research design would, however, have further complicated the research, since it would be desirable to somehow guarantee a degree of similarity between the sessions of different facilitators.

Nevertheless, we hope that we have shown that multicultural aspects are highly relevant for the design of usability evaluation research. Apart from the many methodological complexities mentioned above, there may also be a risk of stereotyping cultural groups, especially when describing experimental conditions and formulating hypotheses. It is important, however, to keep in mind that we are talking about differences between groups, not between individual people. Despite all of the methodological problems that may arise in conducting cross-cultural research, we think this type of research will be essential for technical communicators and researchers to gain a clear view beyond the boundaries of their own culture.

CONCLUSION

The results of this study would be misinterpreted if they were used to criticize the usefulness of usability testing as such. To the contrary, the usability evaluation brought to light many problems in a Web site that may be expected to meet the needs of users all over the world. What the study shows is the need for a better understanding of the many factors that may influence the choice of evaluation methods and their results. Cultural differences between participants are only one of these factors, but as shown by the results of this study, they seem to be of considerable importance. **TC**

REFERENCES

- Arnold, M. 1998. Building a truly World Wide Web: A review of the essentials of international communication. *Technical communication* 45:197-206.

One could argue that the participants' behavior during a usability test reflects a deeper level of the cultural system than the answers they give in a questionnaire.

- Barnum, C. M. 2002. *Usability testing and research*. New York, NY: Longman.
- Becker, S. A. 2002. An exploratory study on Web usability and the internationalization of US e-business. *Journal of electronic commerce research* 3:265–278.
- Brown, P., and S. C. Levinson. 1990. *Politeness: Some universals in language usage*. Cambridge, UK: Cambridge University Press.
- de Jong, M. 1998. *Reader feedback in text design. Validity of the plus-minus method for the pretesting of public information brochures*. Amsterdam, Netherlands: Rodopi.
- de Jong, M., and P. J. Schellens. 1997. Reader-focused text evaluation: An overview of goals and methods. *Journal of business and technical communication* 11: 402–432.
- . 1998. Focus groups or individual interviews? A comparison of text evaluation approaches. *Technical communication* 45:77–88.
- . 2000. Toward a document evaluation methodology: What does research tell us about the validity and reliability of methods? *IEEE transactions on professional communication* 43:242–260.
- . 2001. Readers' background characteristics and their feedback on documents: The influence of gender and educational level on evaluation results. *Journal of technical writing and communication* 31:267–281.
- Diamantopoulos, A., N. Reynolds, and B. Schlegelmilch. 1994. Pretesting in questionnaire design: The impact of participant characteristics on error detection. *Journal of marketing research* 36:295–311.
- Dumas, J. C., and J. S. Redish. 1999. *A practical guide to usability testing*. Revised ed. Exeter, UK: Intellect.
- Hall, E. T. 1977. *Beyond culture*. Garden City, NY: Anchor Press/Doubleday.
- Hoc, J. M., and J. Leplat. 1983. Evaluation of different modalities of verbalisation in a sorting task. *International journal of man-machine studies* 18:283–306.
- Hofstede, G. 1994. *Culture and organizations: Software of the mind*. London, UK: Harper Collins.
- . 2001. *Culture's consequences: Comparing values, behaviours, institutions, and organizations across nations*. 2nd ed. Beverly Hills, CA: Sage.
- Hoft, N. 1995. *International technical communication: How to export information about high technology*. New York, NY: John Wiley.
- La Ferle, C., S. M. Edwards, and Y. Mizuno. 2002. Internet diffusion in Japan: Cultural considerations. *Journal of advertising research* 42(2):65–79.
- Lentz, L., and J. Hulst. 2000. Babel in document design: The evaluation of multilingual texts. *IEEE transactions on professional communication* 43:313–322.
- Marcus, A., and E. W. Gould. 2000. Cultural dimensions and global web user-interface design: What? So what? Now what? *Proceedings of the 6th Conference on Human Factors and the Web*. http://www.tri.sbc.com/hfweb/marcus/hfweb00_marcus.html.
- Middleton, K. L., and J. L. Jones. 2000. Socially desirable response sets: The impact of country culture. *Psychology and marketing* 17:149–163.
- Nielsen, J. 1993. *Usability engineering*. Boston, MA: Academic Press.
- . 2000. *Designing Web usability: The practice of simplicity*. Indianapolis, IN: New Riders.
- Okazaki, S., and J. A. Rivas. 2002. A content analysis of multinationals' Web communication strategies: Cross-cultural research framework and pre-testing. *Internet research* 12:380–390.
- O'Keefe, R. M., M. Cole, P. Y. K. Chau, A. Massey, M. Montoya-Weiss, and M. Perry. 2000. From the user interface to the consumer interface: Results from a global experiment. *International journal of human-computer studies* 53:611–628.
- Peng, K., R. E. Nisbett, and N. Y. C. Wong. 1997. Validity problems comparing values across cultures and possible solutions. *Psychological methods* 2:329–344.
- Rubin, J. 1994. *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York, NY: John Wiley.
- Schrifer, K. A. 1989. Evaluating text quality: The continuum from text-focused to reader-focused

methods. *IEEE transactions on professional communication* 32:238–255.

———. 1997. *Dynamics in document design: Creating text for readers*. New York, NY: John Wiley.

Schweibenz, W., and F. Thissen. 2003. *Qualität im Web: Benutzer-freundliche Webseiten durch Usability Evaluation*. Berlin, Germany: Springer.

Sienot, M. 1997. Pretesting Web sites. A comparison between the plus-minus method and the think aloud method for the World Wide Web. *Journal of business and technical communication* 11:469–482.

Simon, J. S. 2001. The impact of culture and gender on Web sites. *Data base for advances in information systems* 32: 18–37.

Smith, P. B., and M. H. Bond. 1998. *Social psychology across cultures*. 2nd ed. London, UK: Prentice Hall.

Ting-Toomey, S. 1998. Intercultural conflicts styles. A face-negotiation theory. In *Theories in intercultural communication*, ed. Y. Y. Kim and W. B. Gudykunst. Newbury Park, CA: Sage, pp. 213–235.

Trompenaars, F., and C. Hampden-Turner. 1998. *Riding the waves of culture: Understanding cultural diversity in global business*. New York, NY: McGraw-Hill.

van den Haak, M. J., M. D. T. de Jong, and P. J. Schellens. 2003. Retrospective versus concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour and information technology* 22:339–351.

———. Forthcoming. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with computers*.

van Someren, M. W., Y. F. Barnard and J. A. C. Sandberg. 1994. *The think aloud method: A practical guide to modelling cognitive processes*. London, UK: Academic Press.

van Waes, L. 2000. Thinking aloud as a method for testing the usability of Web sites: The influence of task variation on the evaluation of hypertext. *IEEE transactions on professional communication* 43:279–291.

Vatrapu, R. 2002. Culture and international usability testing: The effects of culture in interviews. Master's thesis, Virginia Polytechnic Institute and State University. http://scholar.lib.vt.edu/theses/available/etd-09132002-083026/unrestricted/Vatrapu_Thesis.pdf

MARINDA HALL is a master's student of communication studies at the University of Twente, Netherlands. She is currently working as a qualitative researcher and usability specialist at Interview-NSS, Amsterdam. Contact information: hallmarinda@hotmail.com.

MENNO DE JONG is an associate professor of communication studies at the University of Twente, Netherlands. His main research interest concerns the methodology of applied communication research. He has published many articles about document and Web site evaluation and usability testing, and is currently working on an additional research line focusing on applied research methods in organizational and corporate communication. Contact information: m.d.t.dejong@utwente.nl.

MICHAËL STEEHOUDER holds the chair of technical communication and is head of the Department of Communication Studies at the University of Twente, Netherlands. His research focuses mainly on the design of technical instructions. He has co-authored books on communication skills, forms design, and software manuals, and has published more than 100 articles in Dutch and international journals. He is associate editor of *IEEE transactions on professional communication*. Contact information: m.f.steehouder@utwente.nl.