Journal of Usability Studies
Vol. 13, Issue 4, August 2018 pp. 177–195

Are Two Pairs of Eyes Better Than One? A Comparison of Concurrent Think-Aloud and Co-Participation Methods in Usability Testing

Obead Alhadreti
Assistant Professor
Computing College
Umm Al-Qura University
Al-Qunfoudah
Saudi Arabia
oghadreti@uqu.edu.sa

Pam Mayhew
Senior lecturer
School of Computing Sciences
University of East Anglia
Norwich, Norfolk, UK
P.Mayhew@uea.ac.uk

Abstract

This paper presents the results of a study that aimed to compare the traditional concurrent think-aloud protocol with the co-participation method to determine the benefit of adding an additional participant to the testing session. The two methods were compared through an evaluation of a library website, and their relative validity and utility were measured using four points of comparison: overall task performance, test participants' experiences, quantity and quality of problems discovered, and the cost of employing each method. The results of the study show significant differences between the performances of the two types of testing methods. The co-participation method was evaluated more positively by users and led to the detection of a greater number of minor usability problems. This method, however, was found to require a greater investment of time and effort on the part of the evaluator in comparison to the classical method. This study found no difference between the methods in terms of task performance.

Keywords

usability testing, user studies, user experiences, think-aloud protocols, co-participation, human-computer interaction

Introduction

Usability has increasingly become a major contributor to the success of software systems. For web-based systems, usability is especially critical given that the web user population is expanding in age, expectations, information needs, tasks, and user abilities. Nielsen (1999, p. 9) puts this very succinctly: "The web is the ultimate customer-empowering environment. He or she who clicks the mouse gets to decide everything. It is so easy to go elsewhere; all the competitors in the world are but a mouse click away." In other words, if websites are not sufficiently usable, users will simply abandon them in favor of alternatives that better cater to their needs. Despite the general recognition of the importance of usability for web-based systems, it has been argued that many websites today still fail the most basic tests of usability (Choudrie, Ghinea, & Songonuga, 2013). Appropriate website design and effective evaluation methods can therefore help to ensure that websites are usable.

One of the most widely used methods of evaluating the usability of websites is the concurrent think-aloud protocol (CTA), which requires participants to verbalize their actions and thought processes while interacting with a system (Ericsson & Simon, 1993). This is intended to give evaluators direct insight into the cognitive processes employed by users as they work with an interface. These insights can then be measured and analyzed, and the data used to improve the product's usability. However, different variations of the method exist among usability professionals today. An increasingly common version of the protocol is the co-participation (CP) method, also known as the constructive interaction or team TA (Jeffrey & Chisnell, 1994), and involves two participants working together to explore the test object and perform tasks. The paired participants are asked to engage in verbalizing as they interact with the system and one another.

This paper presents the findings of our study on the effect of adding an additional participant to the test session by comparing performance of the classic CTA method with the CP method. The rest of this paper is structured as follows: the next section discusses the existing literature focusing on recent studies related to the TA methods in usability testing and states the aims and research questions of the current study; further sections discuss the research method, data analysis, and results of this study; and finally, the paper concludes with a brief discussion of the findings.

Related Work

This section presents an overview of the related literature on TA protocols.

History and Theoretical Background

Despite their increasing use within the context of usability testing, CTA and CP methods were originally developed within a relatively narrow niche in the field of cognitive psychology. John Watson (1920) was the first to report on using thinking aloud as he tried to learn more about the psychology of thinking (Fox, Ericsson, & Best, 2011). Duncker (1945; original German version 1935) was among the first researchers to utilize thinking aloud in empirical studies of mathematical problem solving from 1925 to 1940. Later, the verbal reports produced by TA protocols also began to serve as a basis for discovering how people perform certain activities in many other fields: how they write (Hays & Flower, 1983) or read (Ericsson, 1988); what a translation process looks like (Séguinot, 1996), et cetera. Most of the literature devoted to TA protocols is based, to a larger or smaller extent, on Ericsson and Simon (1980), whose influential work has almost single-handedly validated the use of verbal protocols as research data.

TA methods have been employed in usability testing for more than thirty years since their introduction to the field by Lewis and Mack in 1982 (Lewis, 1982) when the CTA method was used to get insight into the users' mental processes as they learned to use new text processing systems. Studies by Jørgensen (1990) and Wright and Monk (1991) have shown that the CTA method is highly effective for detecting usability problems in user interface design, especially if the designers conduct the usability tests themselves and so get direct feedback from the users. Since then, the CTA method has become the method of choice for many usability practitioners (Kumar, Yammiyavar, & Nielsen, 2008).

The Concurrent Think-Aloud Protocol

An international survey conducted by McDonald, Edwards, and Zhao (2012) showed that 98% of usability professionals had utilized CTA and 89% rated it as the most frequently used approach. CTA is attractive to practitioners for a number of reasons, such as its value in providing insight into the actions and intentions of users, and its ability to capture real-time responses from users during the testing process. Perhaps the main reason for its popularity among usability practitioners, however, is that it is fast and easy to implement (Alhadreti & Mayhew, 2017; McDonald et al., 2012). The critical importance of time and cost in the IT industry often means that usability practitioners must conduct tests according to tight deadlines and with limited resources at their disposal (Nørgaard & Hornbæk, 2008). It follows, then, that the most popular testing method would be one that enables practitioners to carry out usability analyses and deliver their reports in a time- and cost-effective manner.

However, the CTA method is not a method without problems. The first of the problems is that the completeness of the data gathered is questionable. Ericsson and Simon (1993) acknowledged that although the concurrent data can provide sufficient evidence for the accurate sequence of thoughts that participants had while completing the task, the verbal reports are likely to be incomplete because participants are expected to give priority to task solving and may therefore fail to report some thoughts (Ericsson & Fox, 2011; Ericsson & Simon, 1993). Within the context of usability testing, research investigating the relationship between eye movements and TA protocols suggest that verbal reports may indeed be incomplete (Cooke, 2010). The second issue is simply that the process of concurrent verbalization may feel uncomfortable or unnatural, as people do not commonly verbalize their thoughts constantly while working (Nielsen, 1993). The third issue concerns the extent to which the request to talk aloud may interfere with and alter participants' thought processes and task performance and risk skewing the validity of the data. The change in task performance is often referred to as reactivity. Usability studies that have compared CTA with a silent condition alone or a silent condition followed by a retrospective thinking-aloud have had mixed results (e.g., Alshammari, Alhadreti, & Mayhew, 2015; Hertzum, Hansen, & Andersen, 2009; Peute, de Keizer, & Jaspers, 2010; van den Haak, de Jong, & Schellens, 2004). Ericsson and Simon (1993) stated that CTA does not cause any reactivity if a minimal interaction is kept between the evaluator and the participants during the TA process. The evaluator should only issue TA reminders if participants fall silent, and the reminders must be short and non-directive, such as "keep talking," to safeguard against reactivity and evaluator-induced bias.

The Co-Participation Method

Apart from the single-user usability evaluation methods, there are also usability approaches involving multiple users. The most common multiple-user method is one that was originally developed by Miyake in the early 1980s (Miyake, 1982). She asked participants to learn how a sewing machine could make stitches and had them work together in teams of two, hoping to find out to what extent the participants' sharing of knowledge would improve their learning process. Miyake's method, which she labelled constructive interaction, was later adopted into the field of usability testing by O'Malley, Draper, and Riley (1984). The CP method is less frequently employed than the single-user methods, but its popularity seems to be rapidly increasing. One obvious reason for the growing interest in this particular method is the development of systems and tools that support collaborative work, both on-site and at a distance. These programs require evaluation, which, for ecological reasons, needs to be carried out by more than one user. Kahler, Kensing, and Muller (2000), for instance, employed CP to evaluate a tool for sharing document templates and toolbars. Her study involved pairs of participants who sat next to each other at separate workstations performing different but related tasks. Nodder, Williams, and Dubrow (1999) used a similar test procedure to evaluate NetMeeting, a real-time conferencing product. They performed iterative testing and situated pairs of participants in different rooms to account for the long-distance communication that is an important feature of online meetings. The CP method is also considered an effective way of making TA test participants feel like the testing experience is more natural (Nielsen, 1993; van den Haak et al., 2004). Nielsen (1993) further stated that the CP method is especially suited to usability evaluations involving children as it better facilitates children's verbalization than does the classical TA protocol. However, using two people for each test increases the cost of testing and the difficulty of finding a sufficient number of test participants (Als, Jensen, & Skov, 2005).

Other TA Protocols

There is a variety of other TA protocols. Ericsson and Simon (1993) introduced retrospective TA (RTA) as a classical TA method wherein participants are asked to verbalize their thoughts after performing the tasks. This method has received less attention compared to the CTA (McDonald et al., 2012). For some practitioners the RTA method is more natural than the CTA one (McDonald et al., 2012); however, RTA does have some drawbacks. One of these relates to the method's reliance on human memory, which is fallible: With the best of intentions, participants might forget specific things that occurred during a task. Ericsson and Simon (1993) stated that some information may be lost in the case of retrospective research, which was confirmed by Peute et al. (2010) and Alshammari, Alhadreti, and Mayhew (2015). Another drawback is that it requires a longer time to apply compared to the CTA method (Alhadreti, 2016).

Based on field observations of usability evaluations, Boren and Ramey (2000) conducted observations of usability practitioners in software companies. They found several discrepancies between the traditional CTA protocol, which requires a minimal interaction between evaluator and participants during the test, and the way the practitioners conducted the evaluations and in particular the probes they used. Boren and Ramey (2000) suggested that a TA protocol based on speech communication theory, referred to here as the speech-communication method, may be better suited to usability research. Boren and Ramey (2000) stated that for usability studies, the traditional TA protocol where the test evaluator remains silent outside of short assertive commands to "keep talking" might be more disruptive to the participant than previously acknowledged, because humans communicate within a speaker/listener relationship. They argued that their protocol reflects the way human beings naturally communicate, with a combination of statements offered by a speaker followed by feedback or acknowledgment from a listener. According to speech communication theory, during a conversation, it is essential for the listener to use verbalized sounds or phrases which affirm to the speaker that the listener is paying attention and is absorbed in the communication act (Boren & Ramey, 2000). Although the speech-communication protocol was designed with usability evaluation in mind, there is no definitive evidence regarding its real contribution, as few research studies have examined it in detail.

Prior Comparison of Co-participation and Single-Participant Methods

There have been few comparative studies that have measured the validity and utility of the CP protocol against that of the traditional CTA protocol. To mention some, Adebesin, De Villiers, and Ssemugabi (2009) compared the CP protocol with the CTA and analyzed the effect of the CP method on task performance. They found no significant differences between the methods. Similar results were found by Als et al. (2005) who also studied the CP and the CTA, and they found that the CP method costs less than the CTA method in terms of the total time expended by the evaluator to conduct testing sessions and to analyze results. They also found that the paired test participants detected significantly higher number of usability problems than did the single test participants. In contrast, van den Haak et al. (2004) found no significant differences between the paired test participants and the single test participants in the number of problems detected or in the task performance measures, but the CP was rated more positively by its users.

Even though the above-mentioned studies have improved the understanding regarding the usefulness of the methods, most of those studies, however, have a serious common drawback in that they failed to control for the "evaluator effect" on the usability problem extraction process. The evaluator effect is defined as the extent to which "multiple evaluators evaluating the same interface with the same usability evaluation method detect markedly different sets of problems" (Hertzum & Jacobsen, 2001, p. 421). This factor might have significant negative consequences on the validity of the comparative study (Alhadreti, 2016). In addition, Adebesin et al. (2009) did not report on the number and kinds of problems detected by the participants in the think-aloud conditions. With problem detection typically being one of the most important functions of usability testing, the researchers thus failed to account for a crucial factor in their comparison of the two methods. Furthermore, in van den Haak et al.'s (2004) study, another important issue was not taken into account: the level of acquaintance between the pairs. Previous studies have indicated that test participants can behave quite differently depending on how well they know each other (Als et al., 2005). These variables, if not accounted for, can

make it difficult to determine cause and effect. The usefulness of the CP method is therefore yet to be examined in detail.

The Present Study

This current study is part of a larger research project that focuses on the merits and restrictions of different variations of TA protocols for website usability testing (Alhadreti, 2016). The first study of the project compared the classical concurrent TA protocol, retrospective TA protocol, and a hybrid method. The second study compared the performance of the classic CTA method with two relaxed variations on this method—namely, the active intervention method and the speech-communication method—and found the CTA to be the most effective method (Alhadreti & Mayhew, 2017). This paper describes the third study which aims to investigate the utility and validity of the CP method by comparing it against the CTA method's performance measured in the second study of the project. The research questions we endeavor to address with this study are as follows:

Research Question 1 (RQ1): Are there differences between CTA and CP methods with regard to participants' task performances?

Research Question 2 (RQ2): Are there differences between CTA and CP methods with regard to participants' testing experiences?

Research Question 3 (RQ3): Are there differences between CTA and CP methods with regard to the quantity and quality of usability problems they detect?

Research Question 4 (RQ4): Are there differences between CTA and CP methods with regard to the cost of employing the methods?

Method

In this section, we describe the methodology used to address the research questions and the strategies considered for analyzing the data. We also describe how usability problems were extracted from the test data and the factors considered to reduce the evaluator effect.

Study Design

To fulfil the aim of the study, we used an experimental approach with a between-group design. The within-group design was rejected because of the possible carryover effects between the TA conditions (Lazar, Feng, & Hochheiser, 2010). The independent variable under examination in this study is the type of TA methods: the CTA and the CP methods. The dependent variables are performance data from participants' tasks, participants' testing experience, usability problem data, and the cost of employing methods.

Test Object and Tasks

While empirical evidence on the effect of TA methods on the usability testing of websites has been limited (Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010), this dearth of evidence is more visible with regard to academic library websites. Therefore, we decided to target a university library website. The website (i.e., the Durham University library website) and task set used in Alhadreti and Mayhew (2017) were the same ones targeted in this study. There were a number of factors supporting this decision. Firstly, this study is directly linked to one of the previous experiment's conditions (CTA condition), so to ensure valid comparison, the same test object must be used. Secondly, there had been no changes to the website design; the first author made an inspection to confirm that the identified problems were still present in the website and contacted the administrator to confirm that there were to be no modifications in the website's design for the duration of the study. Thirdly, the time between these two experiments was short; it did not exceed three months. Nine tasks were used that together covered the targeted website's main features and predicted problematic areas: finding borrowing information (Task 1), finding information regarding off-campus services (Task 2), booking a study room (Task 3), searching the library catalogue using its simple search (Task 4), and searching the library catalogue using the advanced search (Tasks 5-9). All tasks were designed to be carried out independently from one another, meaning that even if a task was not completed successfully, participants could still carry out the other tasks. We conducted a small pilot study with three people to test the tasks prior to the commencement of data collection.

Participants

The recruitment criteria for this study were the same as the ones applied in our previous study (Alhadreti & Mayhew 2017), wherein 20 participants were recruited for the CTA testing condition. A sample size of 20 for each TA method creates sufficient statistical power to provide a stable estimate (Gray & Salzman, 1998) and is also very likely to produce statistically significant findings (Macefield, 2009). As mentioned in Alhadreti and Mayhew (2017), we decided to select the study sample from amongst university students, as the site administrator deemed them the dominant and most important user group of the tested website. The age range of the recruited participants was 18 to 65 years old; the age was limited to 65 years old to limit the influence of ageing on TA usability testing (Olmsted-Hawala & Bergstrom, 2012; Sonderegger, Schmutz, & Sauer, 2016). An attempt was made to recruit participants with similar characteristics to the participants in the Alhadreti and Mayhew (2017) study to mitigate the impact of individual differences. The sample for the CP condition was recruited through various channels, such as personal emails, posters displayed on the University of East Anglia's (UEA) notice boards, requests on social networking sites, and conversations with personal contacts. In addition, an email was also sent through official channels to students studying in the researchers' university (i.e., the UEA in the UK). The email informed prospective participants that they would be asked to invite a friend to join them in the test session, and that they and their friend would each receive £5 as a token of appreciation for participating in the study. The email also provided a link to the online pre-experiment questionnaires, where prospective participants could provide key demographic details about themselves. Twenty students, from the UEA, who met the study requirements were invited via email to participate in the study. The invited participants to the CP condition were then asked to bring a partner to join them in the session, making a total of forty participants, divided into small teams of two. The students were informed that their partners should have, to some extent, similar characteristics to them in terms of gender, age, Internet experience, and so on. The students were also asked to direct their partners to fill out the pre-experiment questionnaires.

Table 1 illustrates the summary statistics of the demographic characteristics of the participants in the present study. The participants in the CP condition were working in pairs, each with a different role. The "CP actor" column in Table 1 refers to the participants working behind the computer in the CP condition, while the "CP co-actor" column refers to those sitting next to the CP actor. As shown in Table 1, 24 men (60%) and 16 women (40%) participated in the CP experiment, and 60% of the CP participants were aged between 18 and 29, 35% between 30 and 39, and 5% between 40 and 50. All participants were frequent users of the Internet and had not visited the targeted site prior to this study. We believe that the independent participant groups were matched successfully, given that a non-parametric Kruskal-Wallis H test (Kruskal & Wallis, 1952) found no statistically significance difference between the TA conditions in terms of nationality ($\chi2(2) = 0.606$, $p = .739$), gender ($\chi2(2)= .555$, $p = .758$), age ($\chi2(2)= 1.78$, $p = .411$), or Internet use ($\chi2(2)= .284$, $p = .241$). Accordingly, it can be stated that the internal validity of the study is ensured.

Table 1. Summary Statistics of Demographic Characteristics of Participants

Characteristics		CTA (n=20)	CP actor (n=20)	CP co-actor (n=20)	Total (n=60)
Country/region	UK	15	13	13	41
	Western European	5	7	7	19
Gender	Male	13	13	11	37
	Female	7	7	9	23
Age	18–29	11	14	10	35
	30–39	9	4	10	23
	40–50	0	2	0	2
Internet use	Daily	18	14	17	49
	At least once a week	2	6	3	11

Procedure

All evaluation sessions were conducted in the same laboratory in the School of Computing Sciences at UEA. The session began with the evaluator (first author) welcoming the participants to the laboratory, after which they were informed that they were going to be evaluating a library website. Next, every participant was asked to read and sign a consent form. Participants in the CTA condition were instructed to talk aloud while performing the tasks and not to turn to the evaluator for assistance; they were however informed that if they did fall silent for a period the evaluator would tell them to keep thinking aloud[1]. In the CP condition, the paired participants were seated at the computer—one of them sitting in front of it and the other next to it—and were explicitly instructed to work together, in these words: "Even though only one of you can actually control the mouse, you have to perform the tasks as a team by consulting each other and making joint decisions. I also want you to state aloud what you are doing." They were also told not to turn to the evaluator for assistance. The participants in both conditions then engaged in a brief TA practice session using the simple and neutral task of looking up the word "chant" in an online dictionary. On completion of this step, the participants then began the experiment proper. During the testing sessions, the evaluator remained in the same room as the participants and only issued think-aloud reminders if the participants fell silent for 15 seconds. In order to control for variation in computer performance, a single laptop connected to the University's network was used for all experiments. The Morae software (2015) was used to record the computer screens and participants' voices. Once the participants had completed the tasks, each participant was asked to fill in two online post-test questionnaires to provide feedback on the evaluated website (the System Usability Scale questionnaire) and the testing experience (Experience with TA Test questionnaire). Finally, the evaluator thanked the participants for taking part and gave each one of them the promised £5 as a token of appreciation for participating in the study.

Usability Problem Identification

The process of the usability problem identification considered a number of measures in order to reduce the evaluator effect and to increase the reliability and validity of data (Hornbæk, 2010). The process consists of two stages. In Stage One (*Individual problems*), each pair's testing video was reviewed to detect usability problems. Data files were selected using a random number generator to reduce order effect. A clear and explicit usability problem indicator checklist was used at this stage to guide the extraction process. We adopted the checklist developed by Vermeeren, Bouwmeester, Aasman, and de Ridder (2002). Zhao, McDonald, and Edwards (2012) and Alhadreti and Mayhew (2017) adopted this checklist in their TA studies, and they found that the checklist increases the reliability of data collected. Each problem that was discovered in our study was assigned a number (e.g., IUP1) and was recorded in a report in terms of the contexts in which they arose, their descriptions, their impact, their persistence, the

[1] The experimental procedure of the CTA condition was described in Alhadreti and Mayhew (2017). However, for the sake of completeness, we include the description of the procedure here.

current task, and time when it occurred. In Stage Two (*Final problems*), starting with the first pair, individual problems were merged across participants to form a final usability problem if they had similar problem descriptions and contexts. Structured reports were also used at this stage to record detailed information relating to each final problem. Each final problem was assigned a unique number (e.g., FUP1). All previous documents, namely individual problem reports, were attached to this final report.

Results

This section presents the results obtained from the two TA methods employed in the study.

Task Performance

Four indicators were used in this study to measure the task performance in the CP condition and to determine whether the method induces reactivity. These indicators included the number of tasks that were completed successfully, the total amount of time required to complete the tasks, the number of mouse clicks made, and the number of pages visited. An independent t-test found no statistically significant difference between the test groups in any measures of the task performance (see Table 2).

Table 2. Task Performance Measures

	CTA		CP		*p*-value
	Mean	**SD**	**Mean**	**SD**	
Tasks completion rate	5.50	1.59	5.85	1.79	$p = .096$
Time on tasks (minutes)	25.15	3.45	28.10	5.70	$p = .093$
Number of mouse clicks	105.20	22.70	110.00	15.69	$p = .134$
Number of pages browsed	34.80	7.86	39.40	11.03	$p = .280$

Participants' Testing Experiences

This subsection presents participants' satisfaction with the website usability and their experience with the TA test. As the participants in the CP condition were working in pairs, each with a different role (actor/collaborator) that may have influenced their experiences, they were treated as separate subgroups in the analyses of the post-test questionnaire results. The actors, that is, the participants working behind the computer, are referred to as CP actor, while the collaborators, that is, those sitting next to the person working behind the computer, are referred to as CP co-actors.

Participants' satisfaction with the usability of the website

We used the SUS questionnaire to investigate the effects of the variations of TA protocols on participants' satisfaction with the tested websites. Table 3 shows that the participants in the conditions did not find the system usable. A one-way ANOVA test was conducted and indicated that the satisfaction rating did not differ significantly between the conditions.

Table 3. Participants' Satisfaction with the Tested Website

	CTA		CP actor		CP co-actor		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
SUS score	61.60	10.58	54.35	10.90	57.20	7.72	p =.073

Note. SUS score is on a totaled scale of 0 to 100.

The participant experience with the TA Test

The participant experience with the TA test questionnaire was based on previous research (van den Haak et al., 2004), and the purpose of the questionnaire is to understand participants' experiences of the TA testing environment. Participants rated their experiences by indicating the extent to which they agreed or disagreed with a number of statements on a 5-point scale, with a rating of 1 for "strongly disagree" and 5 for "strongly agree," as recommended by Lazar et al. (2010). Table 4 presents the results of participants' ratings in the two TA conditions. To start

with, all participants were asked to assess in what respect(s) their working process during the test differed from their normal working process by estimating how much slower and more focused they felt they were while working on the tasks. As shown in Table 4, the participants in all the conditions felt that their work on tasks was not that different from their normal work. The scores for the two items are fairly neutral, ranking around the middle of the scale, with average scores ranging from 2.10 to 3.00. No significant differences were found between the conditions.

Participants were next asked to indicate whether, and to what extent, they felt that having to TA and/or work together was difficult, unpleasant, tiring, unnatural, and time consuming. A Kruskal Wallis H test and Bonferroni post hoc analyses showed that both the CP actor participants and the CP co-actor participants found working together significantly more natural and pleasant than the participants in the CTA condition did about having to talk aloud concurrently. It might be easy to see why working together would be evaluated more positively by participants: participants can share their workload and they can talk to each other in a much more natural way than if they were required to think aloud concurrently while working alone.

The final part of the questionnaire concerned the presence of the evaluator. Participants were asked to indicate to what degree they found it unpleasant, unnatural, and disturbing to have the evaluator present during the study. Interestingly enough, a Kruskal-Wallis H test and Bonferroni post hoc analyses revealed that the CP co-actor participants found the presence of the evaluator to be significantly more unnatural than did the CTA participants. No such differences arose in other aspects. A possible explanation could be the workload of the participants. The CTA participants and the CP actors had to actively perform tasks and talk aloud, which considerably reduced the amount of attention they could spare for noticing the evaluator. The CP co-actor participants, on the other hand, were only helping their partners perform tasks, which might require less concentration and thus make them more aware of the evaluator's presence.

Table 4. Participants' Experience with the TA Test

	CTA		CP actor		CP co-actor		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Working condition							
Slower than my normal working	2.50	1.19	3.00	1.02	2.80	1.25	$p = .315$
More focused than my normal working	2.70	1.36	2.10	0.85	2.30	0.97	$p = .551$
TA experience							
Difficult	2.10	1.07	1.95	0.75	1.70	0.50	$p = .633$
Unnatural**	2.85	0.44	2.05	0.64	1.90	0.85	$p = .002$
Unpleasant**	2.45	1.14	1.55	0.51	1.30	0.59	$p = .001$
Tiring	2.20	1.00	1.80	0.76	1.60	0.52	$p = .115$
Time-consuming	2.60	1.45	2.45	0.88	2.30	1.09	$p = .922$
Evaluator presence							
Unnatural	1.50	0.93	1.90	0.71	2.05*	0.68	$p = .028$
Disturbing	1.45	1.17	1.40	0.54	1.20	0.32	$p = .410$
Unpleasant	1.25	1.23	1.20	0.39	1.50	0.51	$p = .096$

Note. 5-points scale (1: Strongly disagree to 5: Strongly agree), * $p < 0.05$ significance obtained, ** $p < 0.005$ significance obtained

Usability Problems

This section compares the CTA and CP methods in terms of the number and quality of individual (i.e., problems detected per participant/pair) and final usability problems (i.e., problems detected in each condition) that were extracted from the test sessions.

Individual usability problems

Table 5 presents the number of problems discovered during interaction with the website by each testing method, and it also categorizes all problems according to the way in which they came to light: by observation (i.e., problems detected from observed evidence with no accompanying verbal data), by verbalization (i.e., problems detected from verbal data with no accompanying behavioral evidence), or by a combination of observation and verbalization.

A Mann-Whitney test revealed that the CP method detected significantly more individual problems than did the CTA (see Table 5). One explanation for this could be the fact that the CP condition had two pairs of eyes which might allow them to notice more problems on the interface. Another explanation could be that as the CP condition involves two people, they could both suggest possible ways of carrying out the nine tasks. This collaborative way of working might thus offer more opportunities for the participants to encounter and articulate usability problems. With respect to the manner in which the individual problems were detected, as can be seen from Table 5, a Mann-Whitney test reveals that the CP method detected a significantly higher number of individual problems through a combination of observation and verbalization.

Table 5. Number and Source of Individual Problems Identified

	CTA		CP		*p*-value
	Mean	**SD**	**Mean**	**SD**	
Observed	2.50	2.06	2.35	1.51	$p = .698$
Verbalized	2.20	1.28	1.45	0.76	$p = .149$
Both*	6.60	3.78	10.90	4.37	$p = .002$
Total*	11.30	3.96	14.70	4.61	$p = .013$

* $p < 0.05$ significance obtained

Individual usability problems and severity levels

The individual problems detected were categorized into four types according to their impact on participants' task performance: critical, major, minor, and enhancement (Dumas & Redish, 1999, Zhao et al., 2012), as outlined in Table 6.

Table 6. Coding Scheme for Problem Severity Levels

	Problem severity level	**Definition**
1	Critical	The usability problem prevented the completion of a task.
2	Major	The usability problem caused significant delay or frustration.
3	Minor	The usability problem had minor effect on usability, several seconds of delay and slight frustration.
4	Enhancement	Participants made suggestions or indicated a preference, but the issue did not cause impact on performance.

When assigning severity levels to individual problems, the persistence of each problem, which refers to the number of times the same problem is encountered by test participants, was also taken into consideration (Hertzum, 2006). For example, if the same participant encountered the same problem more than three times, even if each incident only had a minor impact, the individual problem was considered as major due to the aggregation of impact (Nielsen, 1993). A Mann-Whitney test found a significant difference between the CTA and CP methods regarding the number of individual problems whose severity was rated as "minor" or "enhancement." The CP method produced significantly more individual minor and enhancement level problems than did the CTA method (see Table 7).

Table 7. Individual Problem Severity Levels

	CTA		CP		*p*-value
	Mean	**SD**	**Mean**	**SD**	
Critical	3.50	0.94	3.25	1.43	$p = .192$
Major	4.20	1.50	4.55	2.67	$p = 833$
Minor*	3.35	2.45	5.60	2.85	$p = .013$
Enhancement*	0.25	0.55	1.30	0.97	$p = .001$

* $p < 0.05$ significance obtained

Individual usability problem types

To enable an examination of the types of problems that were discovered in the CP condition, two usability experts classified all detected problems into four specific problem types: navigation, layout, content, and functionality (see Table 8)[2]. These types are based on the literature related to the categorization of usability problem of online libraries (van den Haak et al., 2004) and the literature related to the categorization of website usability problems (Tullis & Albert, 2008; Zhao et al., 2012). The inter-coder reliability was computed using Cohen's kappa Barendregt, Bekker, Bouwhuis, & Baauw, 2006). The overall kappa was 0.94, which shows a highly satisfactory level of inter-coder agreement.

Table 8. Coding Scheme for Problem Types

Problem type	**Definition**	**Example**
Navigation	Participants have problems navigating between pages or identifying suitable links for information/functions.	The participant has trouble returning to the home page.
Layout	Participants encounter difficulties due to web elements, display problems, visibility issues, inconsistency, and problematic structure and form design.	The participant feels that the font is too small.
Content	Participants think certain information is unnecessary or is absent. Participants have problems understanding the information including terminology and dialogue.	The participant does not understand the feedback of an error messages.
Functionality	Participants encounter difficulties due to the absence of certain functions or the presence of problematic functions.	The participant expects an option on the "Catalogue" page to specify how many items to load per page.

Table 9 shows the number of different types of individual problems identified in the CTA and CP conditions. A Mann-Whitney test revealed that the CP method produced significantly more individual problems compared to the CTA method relating to layout and content problems.

[2] The problems discovered by the CTA condition were previously classified by two usability experts in Alhadreti and Mayhew (2017).

Table 9. Individual Problem Types

	CTA		CP		*p*-value
	Mean	**SD**	**Mean**	**SD**	
Navigation	4.45	1.57	4.80	2.30	$p = .547$
Layout*	4.00	1.86	6.10	2.90	$p = .046$
Content*	0.65*	0.48	1.30	0.86	$p = .020$
Functionality	2.20	1.07	2.50	1.19	$p = .478$

* $p < 0.05$ significance obtained

Final usability problems

In total, 96 problems were extracted from the test session files of the two conditions. The CP method detected 83 final usability problems in the tested website. The CTA method, as mentioned in Alhadreti and Mayhew (2017), detected 60 problems on the website. Accordingly, the CP outperformed the CTA method with respect to the range of final problems detected. The percentages of unique final problems identified by CTA and CP are 13% and 37% respectively. The participants applying the CTA method did not find 36 problems that were uncovered by the CP method. The participants applying the CP method did not find 13 unique problems that had been uncovered by the CTA method. Both groups commonly identified 47 (49%) of the total number of problems.

Final usability problems and their sources

The final usability problems were coded according to verbalization source, observation source, and a combination of both. A problem was deemed to have a combined source if the individual problems had been merged from both verbal and observation sources. To qualify as having either a verbal or observed source, a final problem had to consist of individual problems from a single source of origin (all verbal or all observed; Zhao et al., 2012). The results are shown in Table 10.

Table 10. Final Usability Problem Sources

	CTA			CP		
	Unique	**Overlapping**	**Total**	**Unique**	**Overlapping**	**Total**
Observed	1	6	7	0	5	5
Verbalized	9	8	17	8	3	11
Both	3	33	36	28	39	67
Total	13	47	60	36	47	83

As can be seen from Table 10, the CTA method detected 7 problems derived from observation evidence, 17 from verbal evidence, and 36 from a combination of the two. In the CP test, 5 problems were derived from observation evidence, 11 from verbal evidence, and 67 from a combination of the two. The CP method detected a larger number of both overlapping and unique problems from the combined sources than did the CTA method.

Final usability problems and severity levels

The assignment of severity levels to final problems must take into account the discrepancies between how a given problem may be experienced by participants; for example, a pair may circumvent a problem very quickly, while another may spend a long time overcoming the same problem. To bypass potential conflict between severity levels, levels were assigned according to the majority of problems (Lindgaard & Chattratichart, 2007). In those cases where the contradictory severity levels emerged with an equal number of participants, assignment took place according to the highest severity level (Ebling & John, 2000). Table 11 presents the number of problems for different severity levels from the two TA conditions. The results show that the two methods managed to identify all four critical problems discovered on the site: 31.66% (19 problems) of the final problems from the CTA method were high impact problems

(with critical and major effects), and 68.33% were low impact problems (with minor and enhancement effects), whereas, for the CP condition, 18% (15 problems) of final problems were high impact. In terms of the unique problems, the results revealed that that 38% (5 problems) of the unique problems identified by the CTA method were high impact problems. However, of the problems identified by the CP method, 9% (3 problems) were high impact problems.

Table 11. Final Usability Problem Severity Levels

	CTA			CP		
	Unique	**Overlapping**	**Total**	**Unique**	**Overlapping**	**Total**
Critical	0	4	4	0	4	4
Major	5	10	15	3	8	11
Minor	7	31	38	25	33	58
Enhancement	1	2	3	8	2	10
Total	13	47	60	36	47	83

Final usability problem types

The 96 final problems discovered on the tested website in this study were classified by the usability experts into 23 navigational problems, 42 layout problems, 16 content problems, and 15 functional problems. Table 12 shows the number of final usability problems by their type. Compared with the CTA method, the CP method identified more problems of each type and also detected more unique problems of each type than did the CTA method.

Table 12. Final Usability Problem Types

	CTA			CP		
	Unique	**Overlapping**	**Total**	**Unique**	**Overlapping**	**Total**
Navigation	3	15	18	5	15	20
Layout	5	20	25	17	20	37
Content	4	5	9	7	5	12
Functionality	1	7	8	7	7	14
Total	13	47	60	36	47	83

Reliability of problem identification and classification

As in Alhadreti and Mayhew (2017), an independent evaluator was recruited to carry out an inter-coder reliability check on usability problem analysis. The evaluator independently analyzed two randomly selected testing videos from the CP condition. The any-two agreement formula provided by Hertzum and Jacobsen (2003) was used to calculate the inter-coder reliability across the four videos.

$$Any - two\ agreement = \frac{|Pi\ \cap\ Pj|}{|Pi\ \cup\ Pj\,|}$$

In this equation, Pi and Pj are the problems identified by evaluators "i" and "j" respectively. Its value ranges from 0% in the case of no agreement amongst the evaluators to 100% in the case of full agreement. The average any-two agreement for the individual problem identification across the two videos was 73% (individual agreements were 73% and 72%). The any-two agreement for the final usability problems was 78%. The any-two agreement for the individual and final problems for the CTA was 72% and 75% respectively, as shown in Alhadreti and Mayhew (2017). The reliability of the coding of the problem source and severity level for the CP condition was examined using Cohen's Kappa (Field, 2009). For the individual problem levels, the resulting Kappa value for the problem source was 0.689 and for problem severity it was

0.752. For the final usability problems, the resulting Kappa value for problem source was 0.744, and the severity level was 0.832. This indicates high reliability for the coding.

Comparative Cost

The cost of employing the two TA methods under study was measured by recording the time the evaluator spent conducting testing and analyzing the results for each method. Testing time, recorded via an observation sheet, refers to the time taken to carry out the entire testing sessions, including the instruction of participants, data collection, and solving problems that may arise during test sessions. Analysis time, collected via a web-based free time tracking software called Toggle[3] (Version 2013), refers to the time taken to extract the usability problems from each method's testing data. As is shown in Table 13, the CP method required longer session time (802 minutes) than the CTA method (723 minutes). The total time taken to apply the two verbalization methods was 1,525 minutes. An independent t-test found no significant difference between the conditions with regard to session time (p = 0.096).

Table 13. Temporal Cost

	CTA	CP	Total
Session time	723	802	1,525
Analysis time	865	1,006	1,871
Total time	1,588	1,808	3,396

Note. Time is in minutes.

The total time taken to identify usability problems using the two methods was 1,871 minutes, with the CP method requiring higher amount of time (1,006 minutes) in comparison to the CTA (865 minutes). An independent t-test was showed that the analysis time in the CTA was significantly shorter than in the CP condition (p =.003). The overall results showed that the CTA method incurred shorter time (1,588 minutes) than the CP method (1,808 minutes).

Discussion

This section discusses the study's findings and compares them to some of the related literature. It also discusses the limitations of the study.

TA Methods and Task Performance

The CP method did not have an impact on the participants' task solving process, as the CTA and CP methods show no statistically significant differences in task solving accuracy, efficiency, or navigational behavior. Reactivity, therefore, was not evident in the CP method. The CP participants performed their tasks neither better nor worse than the participants in the CTA condition. This corresponds to earlier findings by Adebesin et al. (2009), Als et al. (2005), and van den Haak et al. (2004). This finding implies that practitioners have a free choice between using the traditional TA method or the CP methods if interested in measuring participant task performance.

TA Methods and Participants' Testing Experiences

The CP method seemed to elicit more positive responses from the participants than the CTA method. This finding seems to be in line with van den Haak et al. (2004) who suggested that interaction between participants during the usability testing session could make the participants feel more comfortable and secure, therefore making them more likely to put forward their opinions. However, despite participants in the main preferring the CP method, the CP collaborators also reported that the presence of the evaluator during testing was more unnatural. This suggests it might be better for evaluators to monitor the CP test from a different room.

Regarding participants' satisfaction with the tested website, the CP method seems to have no distinguishable effect when compared to the classical CTA test. This result indicates that it is legitimate to collect data regarding participants' satisfaction when using co-participation testing.

[3] https://toggl.com/

TA Methods and Usability Problems

The results illustrated significant differences between classical TA and co-participation on the identification of usability problems. The current experiment showed that paired participants found more usability issues than single test participants at both the individual and final problem levels. On average the pairs detected 14 usability problems over nine tasks, whereas the single participants found an average of 11 usability issues for the same number of tasks. It was also found that the CP method identified more low severity problems relating to layout and content problems. These findings concur with Als et al. (2005) who found that paired test participants detected a significantly higher number of usability problems than did single test participants. However, it contradicts van den Haak et al. (2004) who found no such difference. This may be because in the van den Haak study, the researchers did not consider the level of acquaintance between the pairs. In addition, the researchers did not apply a structured approach to extracting the usability problems in order to enhance the validity of the data and safeguard against the evaluator effect.

TA Methods and Comparative Cost

The findings of this study reveal that the CTA method costs less than the CP method in terms of the total time expended by the evaluator to conduct testing sessions and analyze results. This finding contradicts with Als et al. (2005) who found that the CP require less time from the evaluator than the CTA to conduct the tests and analyze the results.

Limitations and Directions for Future Research

As with any research, this study has a number of inevitable limitations that could be improved in future work. First, the compared TA methods in the current study were only applied to a university online library. Evaluating different websites with different types of users, such as websites aimed at elderly people, may yield results that are different from the ones presented in this study. Second, there was a difference of three months between the two experiments, which could produce a cohort effect. However, we attempted to control the environment and equipment to reduce the chance of bias occurring due to participants having different equipment or surroundings. Third, some of the nonsignificant findings have p values less than 0.10, which suggests that had our sample size been larger we would have had more power and would have rejected the null hypothesis. With greater statistical power, the failure to observe a difference would be more meaningful. Lastly, this study did not compare the TA methods to silent working. This is a limitation, however, there are studies that have compared both the classical and relaxed TA to silent working and found that the traditional TA does not lead to reactivity (e.g., Alshammari, Alhadreti, & Mayhew, 2015; Hertzum et al., 2009); we followed this assumption in our work and therefore focused only on comparing the two TA methods.

This study provided some interesting insights, but is certainly not the "final word" on usability testing methods. As mentioned earlier, all participants in the study tested the same type of website, an online university library. With this in mind, it would be useful to replicate the study with different testing interfaces to see if the effectiveness of a method can vary according to these factors. Another suggestion for future research concerns the CP method. It would be of interest to compare different team compositions, such as teams of participants who are acquainted with each other versus teams of participants who have never met before, or mixed gender teams versus all-male or all-female teams. Additionally, as we have seen, the results of the CP study show that the co-participants found the presence of the evaluator unnatural. It would be interesting to experiment with the role of the evaluator during co-participation testing—for example, by comparing the results of a test in which the evaluator remains in the test room with another in which the evaluator monitors the test from afar.

Conclusion

This study has provided a more holistic view than what is currently available in the literature on the validity and utility of CTA versus CP usability testing methods. This was achieved by taking a broader, comparative focus, considering various issues and measures. The results of the study show significant differences between the performances of the two types of testing methods. The co-participation method was evaluated more positively by participants, led to the detection of more minor usability problems, and performed better in terms of the relationship between the sample size and the number of problems detected. The method, however, was found to require

a greater investment of time and effort on the part of the evaluator in comparison to the classical method. This study found no difference between the methods in terms of task performance.

Based on the above findings, it can be concluded that the co-participation method seems to be an appropriate method for those usability practitioners who seek to find a high quantity of problems at low severity levels or feel that it is vital that the participants in their usability test experience their participation as pleasantly as possible. Otherwise the classical method seems to be a more cost-effective method as it has the same ability to reveal high-severity problems, requires less time and effort from the evaluator, and involves rewarding one participant per test session instead of two.

Tips for Usability Practitioners

Having discussed the degree of validity and utility of the two TA methods in the previous sections, we offer the following tips for usability practitioners:

- Practitioners can choose either the traditional CTA or the CP method if they wish to capture user performance in the "real context of use," as these methods do not show any effect on task performance.
- Consider using CP when it is vital that the participants in their usability test experience gauge their participation as being as pleasant and natural as possible.
- For CP tests, the evaluator should be located in a separate monitoring room in order to ensure the ecological validity of the test. Based on the questionnaire data, it was obvious that the CP helpers found the presence of the evaluator unnatural.
- Practitioners who are interested in detecting as many problems as possible, regardless of the quality of these problems, may wish to opt for the CP variant.
- Consider using the CP method when interested in finding higher numbers of low severity usability problems—particularly those relating to layout.
- Another practical aspect that usability testers should take into account when planning to conduct CP tests is that the method require a longer time for the application and analysis of the results than the classic CTA method.

Acknowledgements

The authors would like to thank all those people who took time to take part in the experiments. Thanks also to the anonymous reviewers for their helpful comments.

References

Adebesin, T. F., De Villiers, M. R., & Ssemugabi, S. (2009). Usability testing of e–learning: An approach incorporating co–discovery and think–aloud. In J. McNeill, & S. Bangay (Eds.), *SACLA '09: Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association* (pp. 6–15). New York, NY: ACM.

Als, B. S., Jensen, J. J., & Skov, M. B. (2005, June). Comparison of think–aloud and constructive interaction in usability testing with children. In M. Eisenberg, & A. Eisenberg (Eds.), *Proceedings of the 2005 Conference on Interaction Design and Children. IDC 2005. Boulder, Colorado* (pp. 9–16). New York, NY: ACM.

Alhadreti, O. (2016). *Thinking about thinking aloud: An investigation of think-aloud methods in usability testing* (Doctoral thesis, University of East Anglia, Norwich, Norfolk, UK). Available at https://ueaeprints.uea.ac.uk/61487/

Alhadreti, O., & Mayhew, P. (2017). To intervene or not to intervene: An investigation of three think-aloud protocols in usability testing. *Journal of Usability Studies, 12*(3), 111–132.

Alshammari, T., Alhadreti, O., & Mayhew, P. (2015). When to ask participants to think aloud: A comparative study of concurrent and retrospective think-aloud methods. *International Journal of Human Computer Interaction, 6*(3), 48–64.

Barendregt, W., Bekker, M. M., Bouwhuis, D. G., & Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and after some practice*. International Journal of Human-Computer Studies, 64* (9), 830–846.

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication, 43*(3), 261–278.

Choudrie, J., Ghinea, G., & Songonuga, V. N. (2013). Silver surfers, e–government and the digital divide: An exploratory study of UK local authority websites and older citizens. *Interacting with Computers, 25*(6), 417–442.

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication, 53* (3), 202–215.

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Intellect Books.

Duncker, K. (1945). On problem-solving (L. S. Lees, translator). *Psychological Monographs, 58*(5), i-113. http://dx.doi.org/10.1037/h0093599.

Ebling, M. R., & John, B. E. (2000). On the contributions of different empirical data in usability testing. In *Proceedings of the 3rd conference on Designing interactive systems: Processes, practices, methods, and techniques* (pp. 289–296). ACM.

Ericsson, K. A. (1988). Concurrent verbal reports on text comprehension: A review. *Text—Interdisciplinary Journal for the Study of Discourse, 8*(4), 295–325.

Ericsson. K. A., & Fox. M.e. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011). *Psychological Bulletin, 137* (2), 351–354.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review, 87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1993) *Protocol analysis: Verbal reports as data* (Revised ed.) Cambridge, MA: MIT Press.

Field, A. (2009). *Discovering statistics using SPSS* (2nd ed.). London, UK: Sage.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological bulletin, 137*(2), 316.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-computer interaction, 13*(3), 203–261.

Hays, J. R., & Flower, L. S. (1983). Uncovering cognitive processes in writing: An introduction to protocol analysis. In P. Mosenthal, L. Tamor, & S. A. Wamsley (Eds.), *Research on Writing. Principles and Methods*. New York: Longman.

Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human–Computer Interaction, 15*(1), 183–204.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology, 28*(2), 165–181.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human–Computer Interaction, 13*(4), 421–443.

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

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour and Information Technology, 29*(1), 97–111.

Kahler, H., Kensing, F., & Muller, M. (2000). Methods & tools: Constructive interaction and collaborative work: Introducing a method for testing collaborative systems. *Interactions, 7*(3), 27–34.

Kruskal, W. H., Wallis, W. A. (1952). Use of ranks in one–criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621.

Kumar, J., Yammiyavar, P., & Nielsen, J. (2008). Mind Tape technique–a usability evaluation method for tracing cognitive processes in cross cultural settings. *e–Minds, 1*, 69–85.

Jeffrey, R., & Chisnell, D. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests.* Indianapolis, IN, USA: Wiley Publishing, Inc.

Jørgensen, A. H. (1990). Thinking–aloud in user interface design: A method promoting cognitive ergonomics. *Ergonomics, 33*(4), 501–507.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human–computer interaction*. Hoboken: Wiley.

Lewis, C. (1982). Using the" thinking-aloud" method in cognitive interface design. IBM TJ Watson Research Center.

Lindgaard, G., & Chattratichart, J. (2007, April). Usability testing: What have we overlooked? In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 1415–1424). ACM.

Macefield, R. (2009). How to specify the participant group size for usability studies: A practitioner's guide. *Journal of Usability Studies, 5*(1), 34–45.

McDonald, S., Edwards, H., & Zhao, T. (2012) Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication, 55*(1), 1–17.

Miyake, N. (1982). Constructive interaction. Technical report 113. Center for Human Information Processing. San Diego: University of California.

Nielsen J. (1993). *Usability engineering*. San Francisco, CA: Morgan Kaufmann Publishers Inc. ISBN: 0-12-518406-9.

Nielsen, J. (1999). *Designing web usability: The practice of simplicity*. San Francisco, CA: New Riders Publishing.

Nodder, C., Williams, G., & Dubrow, D. (1999). Evaluating the usability of an evolving collaborative product- changes in user type, tasks and evaluation methods over time. In *Proceedings of the international ACM SIGGROUP Conference on Supporting group wor*k (pp. 150–159). New York, NY: ACM.

Nørgaard, M., & Hornbæk, K. (2008). *Working together to improve usability: Challenges and best practices*. Copenhagen University Technical Report.

Peute, L. W., de Keizer, N. F., & Jaspers, M. W. M. (2010). Cognitive evaluation of a physician data query tool for a national ICU registry: Comparing two think aloud variants and their application in redesign. *Studies in Health Technology and Informatics, 160*(1), 309–313.

Olmsted-Hawala, E. L., & Bergstrom, J. R. (2012). Think-aloud protocols: Does age make a difference. *Proceedings of Society for Technical Communication (STC) Summit, Chicago, IL*.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2381–2390). ACM

O'Malley, C., Draper, S., & Riley, M. (1984). Constructive interaction: A method for studying user-computer-user interaction. In B. Shackel (Ed.) *IFIP INTERACT'84 First International Conference on Human-Computer Interaction* (pp. 269–274).

Séguinot, C. (1996). Some thoughts about think-aloud protocols. *Target, 8*(1), 75–95.

Sonderegger, A., Schmutz, S., & Sauer, J. (2016). The influence of age in usability testing. *Applied Ergonomics, 52*, 291–300.

Tullis, T., & Albert, B. (2008). *Measuring the user experience*. Burlington, MA: Elsevier Inc.

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

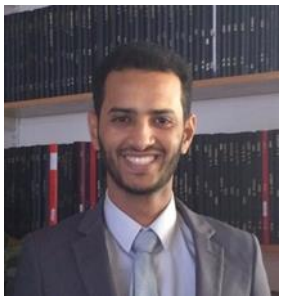
Vermeeren, A. P. O. S., Bouwmeester, K., Aasman, J., & de Ridder, H. (2002). DEVAN: A tool for detailed video analysis of user test data. *Behaviour & Information Technology, 21*(6), 403–423.

Watson, J. B. (1920). Is thinking merely the action of language mechanisms? *British Journal of Psychology, 11*, 87–104.

Wright, P. C., & Monk, A. F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man–Machine Studies, 35*(6), 891–912.

Zhao, T., McDonald, S., & Edwards, H. M. (2012). The impact of two different think-aloud instructions in a usability test: A case of just following orders? *Behaviour & Information Technology, 33*(2), 163–183.

About the Authors



Obead Alhadreti
Dr. Alhadreti is an assistant professor at the Computing College, Umm Al-Qura University, Saudi Arabia. He has been involved in usability testing since 2009. His doctoral research focuses on the use of the TA methods within usability testing. His interests involve usability evaluation, cultural usability, and user experience.

Pam Mayhew
Dr. Mayhew is a senior lecturer in the School of Computing Sciences at the University of East Anglia in Norwich. Her broad interest is in the development of successful, usable systems via appropriate stakeholder participation. This has led to a particular concentration on user centered development, usability testing, and user experience issues.