Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols

ABSTRACT

This paper presents the results of a study that compared three think-aloud methods: concurrent think-aloud, retrospective think-aloud, and a hybrid method. The three methods were compared through an evaluation of a library website, which involved four points of comparison: task performance, participants' experiences, usability problems discovered, and the cost of employing the methods. The results revealed that the concurrent method outperformed both the retrospective and the hybrid methods in facilitating successful usability testing. It detected higher numbers of usability problems than the retrospective method, and produced output comparable to that of the hybrid method. The method received average to positive ratings from its users, and no reactivity was observed. Lastly, this method required much less time on the evaluator's part than did the other two methods, which involved double the testing and analysis time.

Author Keywords

Usability testing; user studies; user experiences; think-aloud protocols; human-computer interaction.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI) --- User Interfaces: Evaluation/methodology, Theory and Methods.

INTRODUCTION

In website design and engineering, the term "usability" describes how easy a website or interface is to use. As the Internet continues to grow exponentially, with millions of websites vying for users' attention, usability has become a critical factor determining whether a website will survive or fail. If websites are not sufficiently usable, users will simply abandon them in favour of alternatives that better cater to their needs [3]. It is therefore crucial that designers employ effective evaluation methods in order to assess usability and improve user interface design. One of the most widely used methods of evaluating the usability of websites is the think-aloud (TA) protocol, wherein users are encouraged to verbalise their experiences, thoughts, actions, and feelings whilst interacting with the interface. This provides direct insight into the cognitive processes employed by users—knowledge which can then inform strategies to improve usability. However, despite the common usage of TA protocol in the field, the specific TA procedures employed vary widely amongst usability professionals [24].

CHI'18, April 21-26, 2018, Montreal, Canada

© 2018 ACM. ISBN

DOI:

The current study investigates the utility and validity of three TA methods, namely concurrent TA, retrospective TA, and a hybrid method, within the context of usability testing. It is part of a larger research project that focuses on the merits and restrictions of different variations of TA protocols for usability testing [1]. The findings of this study will help usability practitioners to make more informed decisions about which TA variant to use in particular contexts.

RELATED WORK

TA methods were originally based on the theoretical framework developed by cognitive psychologists Ericsson and Simon [11], and were introduced to the field of usability testing by Lewis and Rieman in 1982, cited in [20]. According to Ericsson and Simon [12], there are traditionally two basic types of TA methods: the concurrent TA (CTA) method, in which participants TA at the same time as carrying out the experimental tasks; and the retrospective TA (RTA) method, in which participants verbalise their thoughts after they have completed the experimental tasks.

The concurrent method provides "real-time" information during the participant's interaction with a system, which can make it easier to identify the areas of a system that cause problems for the user. This method is the most common TA variant in the field of usability testing [24]. However, there are two main concerns. First, it might be an uncomfortable or unnatural experience, as people do not usually offer running commentaries whilst performing tasks. Second, the request to TA might interfere with and alter participants' thought processes, and may thus affect the ways in which they perform the experimental tasks—which can in turn affect the validity of the data obtained. This change is often referred to as reactivity [38]. By contrast, the retrospective method does not interfere with participants' thought

processes. Participants are therefore fully enabled to execute a task in their own manner and at their own pace, and are therefore less likely to perform better or worse than usual. However, the RTA has been criticised for its reliance on memory, and the subsequent possibility of post-task rationalisations [36].

Ericsson and Simon [13] advocate the use of concurrent and retrospective methods in tandem (referred to as the hybrid (HB) method in this paper). This, they assert, offers a means of enriching the collected verbal data, and of strengthening the validity and reliability of verbal protocols, through the triangulation of concurrent and retrospective data. However, within usability testing, the hybrid method has received very little attention [24]. Indeed, in usability testing research, the concurrent and retrospective TA approaches are typically compared rather than combined [e.g., 23, 36]. At the present time, only a few usability studies have examined the combined use of concurrent and retrospective reporting in the same test. The use of Ericsson and Simon's HB method in usability testing was first investigated by [14]. The results suggested that the interpretation session enhanced the CTA data by adding new problems that were not detected in the CTA phase. A more recent study by McDonald et al. [25] examined the utility of the HB method. The results suggested that the second phase, after the CTA task solving, generated additional insights into the reasons behind the difficulties encountered and decisions made during task performance.

Comparison of Classic Think-Aloud Methods

Ohnemus and Biers [29] were the first to conduct a comparative study of the traditional TA methods. They compared the test participants' performance and subjective ratings in three test conditions: CTA, RTA with reports taken right after the test, and RTA with reports taken on the following day. The results found no significant difference between the groups in terms of either task performance or subjective ratings of the system. Van den Haak et al. [36] conducted a similar study 11 years later, comparing CTA, RTA (with reporting immediately after the test tasks), and the co-participation method. The results showed no significant difference in the total number of problems found, but the problems were detected differently: the retrospective condition revealed more problems through verbalisation, whereas the concurrent condition revealed more problems through observation. Even so, the study found no significant difference in the severity of problems detected, in the participants' overall task performance, or in their experiences with the TA test. Another study by Peute et al. [32] compared the performance of the CTA and RTA, and showed that the CTA method performed significantly better than the RTA in detecting usability problems. In addition, CTA was more thorough in detecting usability problems of a moderate and severe nature. That said, CTA was found to prolong the task processing time.

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, a factor that might have significant negative consequences on the validity of the comparative study [18]. Furthermore, there is a need for a thorough and holistic assessment of the methods. TA protocols have been evaluated based on a range of criteria, including usability problem identification [36], task performance metrics [4], participants' testing experiences [37], and the cost of employing methods [23]. The failure of previous studies to combine evaluation criteria has resulted in conflicting findings and an incomplete understanding. Additionally, no previous study has compared the HB method to any of the one-phase methods (such as CTA or RTA) to truly determine the utility of the approach.

METHOD

Study Design

To fulfil its aim, the study used an experimental approach with a between-group design. The within-group design was rejected because of the possible "carry-over" effects between the TA conditions [20]. The independent variable under examination in this study is the type of TA method: the CTA, the RTA, and the hybrid 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

We decided to use a university library website as a test object for the experiment in this study due to the growing popularity and widespread use of academic digital libraries, and the scant research that investigates the impact of TA methods on usability testing for such media. After a careful evaluation of several websites, the University of East London (UEL) library (UEL-L) website was deemed a promising candidate for this study. This website was chosen because it possessed a certain number of potential usability problems, as determined by a preliminary heuristic evaluation conducted by the first author, and this thereby would ensure to some extent that test participants would encounter difficulties whilst using the site. Once the website was selected, the first author contacted the website administrator via email to obtain consent to use the site, and to establish in advance that there was no intention to modify or alter the interface, either prior to, or during the study.

After defining the test object, a set of tasks was developed to assess the usability of the chosen website by means of the three TA methods. Seven tasks were designed that together covered the targeted website's main features and predicted problematic areas. Task one evaluated the ease of navigating the site. to find the name of a subject support. Task two assessed the booking function for study rooms on the site. Tasks three and four evaluated the site catalogue's "simple search" while tasks five and six evaluated the catalogue's "advaced search" and "sort results" functions. Finally, task seven examined how participants worked with viewing

search history on the site. These tasks were intended to be neither too difficult nor too simple, as both extremes might prevent participants from verbalising and would negatively affect the time required to carry out the tasks [13]. 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. The tasks were piloted with three people prior to the commencement of data collection. An example task is shown below:

> 'Task #4: You want to find the journal paper that has the title "Building for the Future" written by Doyle Henry in 1963 to read before a coming seminar in an education subject. Can you find it?

Participants

The question of what constitutes an optimal number of participants for a usability test is one of the most heated debates in the field. Some researchers state that five to nine participants are sufficient for an effective usability test [26, 27]. However, these numbers are arguably not applicable to the current study, as it aims to investigate the use of different TA usability testing methods rather than to detect usability issues using only a single method. For this study, it was decided that 20 participants would be recruited to each TA testing condition. This figure was based on the grounds that this study is not a typical stand-alone usability test where five to nine subjects are (controversially) adequate, but an experimental study of the relationships between independent and dependent variables which needs more participants to ensure statistical validity [15].

As with tasks, the most important consideration for usability participants is that they are representative of the targeted user groups of the product being evaluated in order to provide the valid feedback needed to make meaningful improvements to a design [34]. To understand the target audience of the system under evaluation, a context of analysis of the tested website was conducted with the website administrator, as recommended by Sova and Nielsen [34]. The site administrator indicated that the library site mainly caters, as expected, for students who are the dominant users of the site (85% of the site's users are students) and academic staff at UEL, although it can also be accessed by other staff and guests (i.e. people outside the university), who together represent its secondary users. We decided to select the study sample from among 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 64 years old; the age was limited to 65 years old to limit the influence of ageing on TA usability testing [31, 33].

Sixty students, from the University of East Anglia (UEA) in the UK, meeting the selection criteria were contacted and invited via email. The sixty volunteers recruited for the study were allocated to the three TA testing conditions, with 20 per condition. To mitigate the impact of individual differences and to be able to draw valid comparisons between the TA groups, participants were matched on the basis of demographic variables as closely as possible. Participants with similar profiles were evenly assigned to the three testing groups in a matched randomised way, using a random number generator.

Table 1 summarizes the demographic profile and descriptive statistics of the participants. All participants were native English speakers, used the Internet on a daily basis and had done so for more than five years, but none of them had ever used the evaluated website or participated in a TA usability test before. Due to having experience with the type of site used as the test object (a university library website) and being part of the target group (university students), but being novice users of the targeted website, the participants were suitable for testing the usability of the UEL-L website. We believe that the independent groups were matched successfully, given that a non-parametric Kruskal-Wallis H test with an alpha level of 0.05 revealed no statistically significant difference between the TA groups in terms of nationality ($\chi2(2)= 2.10$, $p= .34$), gender ($\chi2(2)= .13$, $p= .93$), age ($\chi2(2)= 3.48$, $p= .17$), and or Internet use ($\chi2(2)= .00$, $p= 1.0$). Therefore, the internal validity of the study is high.

Characteristics		CTA	RTA	HB	Total
Country	Britain	18	20	18	56
	Australia	1	0	2	3
	Singapore	1	0	0	1
Gender	Male	11	10	11	32
	Female	9	10	9	28
Age	18-29	15	18	13	48
	30-39	5	2	7	12
Internet use	Daily	20	20	20	60

Table 1. Summary statistics of demographic characteristics of participants

Experimental procedure

All the experiments were conducted in the same laboratory at UEA. When participants arrived at the laboratory, they were cordially greeted by the evaluator (first author) and made to feel at ease. Participants were then asked to review and sign an informed consent form.

HB condition: In the concurrent phase of the HB condition, participants were first asked if they were right- or left-handed (for mouse configuration), and were given a maximum of two minutes to familiarise themselves with the test laptop and to regain their normal speed of interaction with computer systems. On completion of this step, the evaluator introduced the concept of thinking aloud using Ericsson and Simon's

instructions [13]. Participants were instructed to TA while performing the tasks and to not turn to the evaluator for assistance; they were also informed that if they fell silent for a while, the evaluator would remind them to keep thinking aloud. These instructions were followed by a brief TA practice session, as recommended by Ericsson and Simon [13], in which participants were invited to practice thinking aloud using a simple, neutral task of looking up the word "carol" in an online dictionary (unrelated to the use of selected website). After the practice session, the evaluator presented the task instructions sheet to the participants, who were asked to read the instructions first to make sure they understood these fully before proceeding to task solving.

After introducing the test website and setting up the screen capture software (Camtasia), participants began to perform each task in turn. During participants' task performance, the evaluator strictly followed Ericsson and Simon's [13] guidance, and only issued a neutral TA reminder ('please keep talking') if the participants fell silent for 15 seconds; there were no other interactions.

After all tasks were completed, the evaluator ended the recording and directed the participants to fill in the first online post-test questionnaire, the System Usability Scale (SUS) designed by Brooke [8], to assess their satisfaction with the usability level of the tested website. Having done that participants were then asked to complete the first two parts of the second post-experiment questionnaire (Experience with the TA Test), containing questions on their estimation of their method of working on the tasks compared to their normal working (part one), and their experience of thinking aloud (part two) in order to measure their testing experience. This phase was considered complete as soon as participants were finished.

Once the concurrent phase was complete, the evaluator introduced the retrospective phase using Ericsson and Simon's [13] instructions. Participants were asked to watch their recorded performance on muted video and give retrospective reporting. During this phase, the evaluator did not intervene, apart from reminding participants to TA if they stopped verbalising for 15 seconds. Upon completion, the questions posed in the second part of the TA testing experience questionnaire regarding the experience of having to TA were repeated after the retrospective phase in order to investigate whether participants would have different experiences of thinking aloud after the retrospective stage. Afterwards, the participants filled in the third part of the participants' testing experience questionnaire (evaluator presence), including questions on their opinions regarding the presence of the evaluator.

CTA condition: The instructions and procedure for the CTA condition were exactly the same as for the concurrent phase in the HB condition. However, participants in the CTA condition filled in all parts of the post-experiment questionnaires at the very end of the experiment.

RTA condition: In the RTA condition, the evaluator first instructed participants to familiarise themselves with the laptop and perform the preliminary task. They were subsequently asked to review the task instruction sheet and then to solve the seven tasks in silence without the assistance of the evaluator. During testing, the evaluator observed and took notes, but did not interact with participants. At the end of the final task, the participants were asked to fill in the SUS questionnaire, and the first part of the Experience with the TA Test questionnaire. They were then instructed to voice their thoughts retrospectively while watching muted videos of their actions. The instruction for this stage was exactly the same as for the retrospective phase in the HB condition. Subjects were then able to practice thinking aloud. After completing the retrospective reporting, participants were directed to fill in the remaining parts of the Experience with the TA Test questionnaire.

RESULTS

Task Performance

To measure task performance, the number of successful task completions (also known as task success) and the time spent on tasks were collected. The RTA participants in the silent condition were the control group, with results from the other two groups compared against the RTA group's results. By having the CTA and HB groups thinking aloud while performing their tasks, the issue of reactivity would be examined on two fronts. Table 2 shows the results of both indicators. No significance differences were found among the three verbalization conditions in any of the task performance measures. This finding lends support to Ericsson and Simon's [13] argument that thinking aloud does not have an effect on task performance.

	CTA		**RTA**		**HB**		***p*-value**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Task success	4.90	1.34	4.45	0.94	4.75	1.22	.259
Time on tasks (min)	20.67	4.07	18.90	3.76	19.95	3.50	.149

Table 2. Task performance measures

Participants' Experiences

Participants' Satisfaction with the Usability of the Targeted Website

In order to gauge the effect of thinking aloud on participants' perceptions of the usability of the chosen website, participants were asked to fill out the SUS form. SUS scores have a range of 0 to 100, with a higher score reflecting greater participant satisfaction with a site [8]. A one-way ANOVA test indicated that the mean satisfaction scores did not differ between the conditions (see table 3). Apparently, thinking aloud while performing tasks had no effect on participants' satisfaction with the evaluated website.

However, the three participant groups did not find the system very usable.

	CTA		RTA		HB		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
SUS score	70.60	14.73	65.47	17.82	62.55	13.37	.257

On a totaled scale of 1 to 100

Table 3. Participants' satisfaction with the tested website

Participant Experience with the TA Test

The participant experience with the TA test questionnaire was based on previous research [36], and aims to understand participants' experiences of the TA testing environment. Table 4 and 5 present the results of participants' ratings in the three TA conditions. To begin with, all participants were asked to assess how their working procedure on test tasks differed from their usual work approaches by estimating how much slower and how much more focused they were while working on the tasks. As shown in table 4, participants in all three 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, and no significant differences were found between the conditions.

Participants were next asked about the degree to which they felt having to TA (concurrently or/and retrospectively) was difficult, unnatural, unpleasant, tiring, and time-consuming. As shown in table 5, a Kruskal-Wallis H-test and Bonferroni post hoc analyses revealed significant differences between the conditions for "time-consuming". The analysis indicated that the participants in the RTA-HB phase found thinking aloud retrospectively to be more time-consuming than did participants in the CTA-HB phase and participants in the CTA and RTA conditions . This difference may be explained by the longer duration of the HB test and the request for participants to provide dual elicitations, which may have caused the HB participants to rate the TA experience in the retrospective phase as more time- consuming than in the concurrent phase, and as more time-consuming than did participants in the other two conditions. For other items, the participants rated their experiences with thinking aloud as neutral to positive on average. This meant that participants in the CTA and the CTA-HB conditions did not experience reactivity while carrying out tasks.

	CTA		RTA		HB		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Working condition							
Slower	2.40	1.09	2.15	1.30	2.65	1.22	.264
More focused	3.05	1.14	2.80	1.36	3.20	1.70	.638
Evaluator presence							
Unnatural	1.35	0.81	1.80	1.21	1.50	0.88	.302
Disturbing	1.20	0.44	1.60	0.50	1.40	0.51	.378
Unpleasant	1.10	0.30	1.30	0.57	1.25	0.44	.386

Five-points scale (1: Strongly disagree to 5: Strongly agree)

Table 4. Participants' experience with the test

The final part of the Experience with the TA Test questionnaire included measurement items about the presence of the evaluator. Participants were asked to indicate to what extent they found it unnatural, disturbing, and disturbing to have the evaluator present during the study. Kruskal-Wallis H-test testing yielded no significant differences between the conditions regarding these questions (see table 4). As the average scores of the participants ranged between 1.10 and 1.80, the participants clearly felt that the evaluator's presence did not affect their testing experience.

	CTA		RTA		CTA- HB		RTA-HB		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Difficult	2.60	0.88	2.35	1.26	2.50	1.19	2.20	1.32	.304
Unnatural	3.05	0.94	2.75	0.85	3.30	0.80	2.90	1.61	.228
Unpleasant	2.65	1.38	2.40	1.56	2.45	1.14	3.00	1.37	.406
Tiring	2.50	1.19	2.00	0.85	2.30	0.97	2.80	1.36	.282
Time-consuming*	2.70	1.48	3.05	1.30	2.90	1.43	4.25	0.91	.010

Five-points scale (1: Strongly disagree to 5: Strongly agree); * p< 0.05 significance obtained

Table 5. Participants' experience with the TA process

Usability Problems

We considered a number of measures during the process of identifying the usability problem in this study in order to reduce the evaluator effect and to increase the reliability and validity of data [19]. This process is explained in detail in [2]. This subsection presents the results relating to the quantity and quality of usability problem data at the level of individual problems (i.e., problems detected per participant in each condition) and final problems (i.e., the aggregate problems detected in each condition). Since the individual usability problem data were not normally distributed, a Kruskal-Wallis H-test was used to analyse the data. Descriptive data is presented for the final problem set.

The Number of Individual Usability Problems

The most common way to measure usability issues is to count the number of problems found [35]. Table 6 presents the mean number and standard deviation for problems detected per participant, and classifies all problems according to how they were detected: (1) through observation (i.e., from observed evidence with no accompanying verbal data), (2) through verbalization (i.e., from verbal data with no accompanying behavioural evidence), or 3) through a combination of observation and verbalization [36]. As can be seen in table 6, A Kruskal-Wallis H-test revealed and Bonferroni post hoc analyses indicated that the RTA participants discovered significantly fewer individual problems than participants in the CTA and HB conditions. A possible explanation for this discrepancy is that asking test participants to report problems after performing tasks silently may have increased their likelihood of forgetting to report problems during the retrospective phase, even if they had noticed these problems while performing tasks. This finding lends support to Ericsson and Simon's [13] argument that vital information may be lost when applying retrospective research methods, and casts doubt on the validity of the outcome of a RTA evaluation as an overall indication of usability. However, no significant differences were detected between the results of the HB and CTA conditions, suggesting that thinking both concurrently and retrospectively did not cause the HB participants to detect a substantially larger or smaller number of individual problems than the CTA participants. The HB participants not finding a significantly larger number of individual problems may be attributed to their feeling that they had already provided detailed comments in the concurrent phase, and/or feeling tired due to the prolonged duration. The fact that the HB participants did not detect a significantly smaller number of problems than the CTA participants could be attributed to their providing a full account during the concurrent reporting phase, which led them to detect a comparable number of problems to the CTA participants.

Individual Usability Problems and their Sources

With respect to the manner in which the individual problems were detected, it can be seen from table 6 that participants' verbalisations in all three conditions aided them in detecting problems that were not otherwise observed (verbalised problems), or in emphasising or explaining problems that were also observed in their actions (combined problems). This result confirmed the invaluable contribution of verbal protocols to the outcome of usability testing that numerous scholars have highlighted in previous research [e.g., 27, 9, 6].

	CTA		RTA		HB		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Observed	1.35	0.74	1.30	0.47	1.20	0.41	.773
Verbalised*	2.65	1.75	1.00	1.25	2.75	2.48	.004
Both	5.55	1.63	4.05	1.98	5.95	3.82	.071
Total*	9.55	3.26	6.35	3.09	9.90	5.33	0.16

*** p< 0.05 significance obtained**

Table 6. TA methods and the number of individual problems

A Kruskal-Wallis H-test and Bonferroni post hoc analysis showed that the CTA and HB participants detected a significantly higher number of verbalised individual problems than the RTA participants. There were no differences in the number of individual problems detected through evaluator observation or the combined source. However, as the CTA and HB participants did not experience more observable difficulties than the RTA participants, this once again supports Ericsson and Simon's (1993) argument that thinking aloud while performing tasks does not negatively affect performance.

Individual Usability Problems and Severity Levels

The severity levels of individual problems were categorised into one of four types according to their impact on participants' performance: 1) critical, 2) major, 3) minor, and 4) enhancement [9, 2, 38], as outlined in Table 7.

	Problem Severity level	**Definition**
1	Critical	The problem prevented the completion of a task
2	Major	The problem caused significant delay (more than one minute) or frustration
3	Minor	The 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

Table 7. Coding scheme for problem severity levels

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 a test participant, was also taken into consideration [17]. Table 8 presents the mean value and the standard deviation of the number of individual problems at each severity level. A Kruskal-Wallis H-test and a post hoc analysis showed that the CTA and HB

participants found a significantly higher number of minor problems than the RTA participants. There were no significant differences between the methods for the number of individual critical, major or enhancement problems detected.

	CTA		RTA		HB		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Critical	1.90	0.74	2.20	0.83	2.15	0.91	.375
Major	2.90	1.74	2.15	1.84	2.50	2.55	.314
Minor*	4.40	3.74	1.80	1.63	4.65	4.30	.014
Enhancement	0.35	0.48	0.20	0.62	0.60	1.48	.933

*** p< 0.05 significance obtained**

Table 8. TA methods and individual problem severity level

Individual Usability Problem Types

Two independent usability experts were asked to classify the detected problems from the study into four types, as outlined in table 9. These types are based on an initial review of the data, the literature related to the categorisation of usability problem of online libraries [36], and the literature related to the categorisation of website usability problems [35, 38].

Problem type	**Definition**
Navigation	Participants have problems navigating between pages or identifying suitable links for information/functions.
Layout	Participants encounter difficulties due to web elements, display problems, visibility issues, inconsistency, and problematic structure and form design
Content	Participants think certain information is unnecessary or is absent; Participants have problems understanding the information including terminology and dialogue
Functionality	Participants encounter difficulties due to the absence of certain functions or the presence of problematic functions

Table 9. Coding scheme for problem severity levels

Inter-coder reliability was computed using Cohen's kappa [7]. The overall kappa value was 0.87, which shows a highly satisfactory level of inter-coder agreement. The coders discussed the problems that were classified in different categories and created a final classification of all problems on which they both agreed. Table 10 shows the number of different types of individual problems identified in the TA methods. In all conditions, navigation clearly presented the most problems to the participants. This is likely because in working with the tested site, the participants had to navigate many menus of links, each of which they had to interpret before being able to move on to the next level. A Kruskal-Wallis H-test and Bonferroni post hoc analysis showed significant differences between the conditions regarding layout problems: both the CTA and HB participants reported more layout problems than participants in the RTA condition, with the verbalisation conditions bringing to light the other three problem types with similar frequency.

	CTA		RTA		HB		*p*-value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Navigation	4.55	3.42	3.85	3.34	4.90	3.56	.607
Layout*	3.10	2.22	1.00	0.85	3.25	2.20	.002
Content	0.85	0.48	0.60	0.59	0.55	0.60	.164
Functionality	1.05	0.82	0.90	0.44	1.20	1.32	.795

*** p< 0.005 significance obtained**

Table 10. TA methods and individual problem type

The Number of Final Usability Problems

After analysing all of the usability problems found across conditions, the number of problems encountered by all participants were collected, excluding any repeated problems to arrive at a total number of final usability problems. In total, 75 final usability problems were extracted from the test sessions in the three TA conditions. Participants in the CTA condition identified 47 out of the 75 final problems (62%), 13 of which were unique problems. Participants in the RTA condition identified 33 final problems (44%), 8 of which were unique problems, while participants in the HB condition identified 52 final problems, 17 of which were unique problems (see Table 11). Therefore, with respect to the detection of final problems, the CTA and HB methods were again more successful than the RTA method.

Further analysis of the HB condition results revealed that 25 of the 52 total final problems (48%) were detected in the concurrent phase, whereas 5 problems (10%) were only found in the retrospective phase, and 22 problems (42%) were duplicated between both phases, meaning that the majority of the final problems (90%) were in fact detected in the concurrent phase. This reinforces the claim that the retrospective phase has a limited capacity to contribute to usability problem detection, and that the combination of concurrent and retrospective phases advised by Ericsson and Simon [13] may be less beneficial than expected in terms of the quantity of usability problems detected.

Although there were 20 problems (26%) that occurred in all of the three conditions, the overlap between two rather than three conditions was considerably less, ranging from 2% to 16%. These low percentages indicate a substantial number of unique problems identified by three conditions (38 problems). The HB participants discovered twice as many unique problems as the RTA participants. The Venn diagram in Figure 1 shows the overlap between the three conditions.

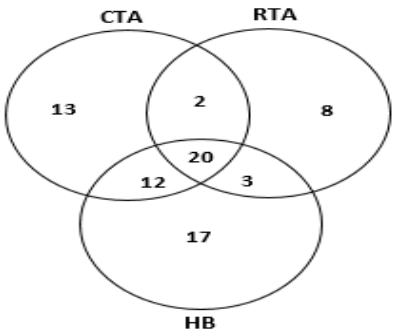

Figure 1. Venn diagram showing overlap in problems between TA protocols

Final Usability Problems and their Sources

Final usability problems were coded according to verbalisation source, observation source, and a combination of both. A problem was deemed to have a combined source if the individual problems had been emerged 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) [38].

As shown in Table 11, the results for the CTA condition were that 6 problems were derived from observation evidence, 15 from verbal evidence and 26 from a combination of the two. In the RTA condition, 7 problems were derived from observation evidence, 6 from verbal evidence and 20 from a combination of the two. In the HB condition, 3 problem were derived from observation evidence, 17 from verbal evidence and 32 from a combination of the two. While the CTA (15 problems) and HB (17 problems) encouraged more verbalised final problems than the RTA (6 problems), a larger number of the unique problems in the CTA (69%), the RTA (62%), and the HB (82%) conditions were derived from verbalisation. With respect to the 5 problems detected in the retrospective phase in the HB condition, all of these were derived from verbalisation.

	CTA		RTA		HB	
	Unique	**Ov.***	**Unique**	**Ov.**	**Unique**	**Ov.**
Observed	0	6	0	7	0	3
Verbalised	9	6	5	1	14	3
Both	4	22	3	17	3	29
Total	13	34	8	25	17	35

*** Overlapping**

Table 11. TA methods and final problem sources

Final Usability Problems and Severity Levels

The assignment of severity levels to final problems took into account the discrepancies between how a given problem may be experienced by participants; for example, one participant 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 [22]. In those cases where the contradictory severity levels emerged with an equal number of participants, assignment took place according to the highest severity level [10].

Table 12 presents the number of problems according to severity level for the three TA conditions. As shown in the table, while the three methods identified the same numbers of critical problems, the distribution of severity differed between each method. 28% (13 problems) of the final problems from the CTA method were high impact problems (with critical and major effects), and 70% (34 problems) were low impact problems (with minor and enhancement effects). For the RTA condition, 39% (13 problems) of final problems were high impact, and for the HB condition, 23% (12 problems) of final problems were high impact. The final five problems found only in the retrospective phase in the HB condition were all minor problems. Regarding unique problems, analysis indicated that no one method identified critical problems that were not identified by the other methods. Analysis also revealed that 15% of the unique problems identified by CTA participants were high impact problems, 25% of the unique problems identified by RTA participants were high impact, and 17% of the unique problems identified by HB participants were high impact.

	CTA		RTA		HB	
	Unique	**Ov.***	**Unique**	**Ov.**	**Unique**	**Ov.**
Critical	0	2	0	2	0	2
Major	2	9	2	9	3	7
Minor	9	21	5	13	12	23
Enhancement	2	2	1	1	2	3
Total	13	34	8	25	17	35

*** Overlapping**

Table 12. TA methods and final problem severity levels

Final Usability Problem Types

Table 13 shows the number of final usability problems for each problem type according to each TA condition. Of the 75 final problems detected, there were 20 navigational problems, 28 layout problems, 14 content problems, and 13 functional problems. CTA and HB participants identified more problems of each type than RTA participants. The distributions of problem types were similar in the CTA and RTA conditions, with the least frequent being content, then functionality, then layout, and finally navigational problems being the most frequent. The HB condition showed a similar pattern, with the exception of layout problems being the most frequent and navigational problems being the second most frequent. In terms of the unique problems found by the three methods, HB participants seemed to detect more unique layout problems than CTA and RTA participants. With regard to the problems generated from the retrospective

phase of the HB condition, three of these were layout problems and two were content problems.

	CTA		RTA		HB	
	Unique	**Ov.***	**Unique**	**Ov.**	**Unique**	**Ov.**
Navigation	3	14	2	10	1	14
Layout	5	10	2	7	8	12
Content	3	3	3	2	5	2
Functionality	2	7	1	6	3	7
Total	13	34	8	25	17	35

* **Overlapping**

Table 13. TA methods and final problem types

Reliability of problem identification and classification

An extra evaluator was recruited to carry out an inter-coder reliability check on usability problem analysis. The independent evaluator analysed six randomly selected testing videos (two from each condition). The any-two agreement formula provided by Hertzum and Jacobsen [18] was used to calculate inter-coder reliability across the six videos:

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

The average any-two agreement for individual problem identification across the six videos was 67% (individual agreements were 70%, 63%, 69%, 74%, 66%, and 58%). The any-two agreement for final usability problem production was 72% (CTA: 70%, RTA: 78%, and HB: 68%). Overall, the agreements are high compared to those set out in Hertzum and Jacobsen's [18] study, wherein agreements between evaluators ranged from 5% to 65%. The reliability of the coding of the problem source and severity level was examined using Cohen's kappa. For individual problems, the kappa value for problem sources was 0.819, and 0.654 for problem severity. For final problems, the kappa value for problem sources was 0.826, and 0.693 for severity. These values reveal a high degree of reliability for the coding.

Comparative Cost

The cost of employing the three TA methods under study was measured by recording the time the evaluator spent conducting testing and analysing the results for each method. Session time, recorded via an observation sheet, refers to the time required to carry out full testing sessions, including the instruction of participants, data collection, and solving any problems that may arise during the session. Analysis time, collected via web-based free time tracking software called "Toggle" (Version 2013), refers to the time required to extract usability problems from each method's testing data.

Table 14 shows the time spent by the evaluator (first author) on applying and analysing the results for the three verbalisation methods. As is clear from the table, the CTA method required the shortest session time (640 minutes), whereas the HB method required the longest session time (1233 minutes). The RTA testing lasted for 1164 minutes. ANOVA testing and Tukey post hoc analysis revealed that RTA and HB session times were significantly longer than CTA session times. No significant difference was found between the RTA and HB conditions. The total time taken to identify usability problems using the three methods was 2964 minutes, with the HB method requiring the most time (1150 minutes) in comparison to the CTA (733 minutes) and RTA methods (1081 minutes). ANOVA testing and a Tukey post hoc analysis were conducted, concluding that analysis time was significantly longer for the HB condition than for the CTA and RTA conditions.

	CTA	**RTA**	**HB**	**Total**
Session time (m)	640	1164	1233	3037
Analysis time (m)	733	1081	1150	2964
Total time (m)	1373	2245	2383	6001

Table 14. TA methods and time expended

Time per problem can be calculated by dividing the time the evaluator spent on a method by the number of problems identified by that method [2]. The CTA method required 29 minutes per usability problem, whereas the RTA method required 68 minutes per usability problem and the HB method required 45 minutes per usability problem. Therefore, based on the results presented, the outcomes and the time and effort required by the evaluator favour CTA testing over RTA and HB testing.

DISCUSSION

Think-Aloud Methods and Participants' Task Performance

Verbalising thoughts while working did not affect participants' task performance; that is, whether or not a participant was asked to TA during a usability session did not lead to a change in their task success rate or time spent on tasks. Reactivity was therefore not evident here. This implies that the task performance data collected when using concurrent thinking aloud can offer an accurate representation of real-world use. If usability practitioners wish to portray user performance in the "real context of use", they can thus choose between the CTA or HB methods on one hand and the RTA method on the other. These findings both correspond with and contradict earlier work by van den Haak et al. [36], who found no differences in task performance between CTA and RTA methods but did find that thinking aloud led to significantly greater task accuracy. One possible explanation for this discrepancy is that van den Haak's et al. [36] study did not take steps to control the participants' individual differences by matching them as closely as possible between conditions, as was done in the current study. Participants' demographic variables may therefore have affected van den Haak et al.'s results.

Think-Aloud Methods and Participants' Experience

With regards to participants' satisfaction with the tested website, thinking aloud while performing tasks seemed to have no effect on the perceived usability of the tested website, as assessed via comparison with participants in the silent RTA condition. This finding indicates that it is valid to collect data regarding participants' satisfaction when using CTA testing, which is in line with the findings of Olmsted-Hawala et al. [30]. As in van den Haak et al. study [36], the CTA and RTA participants in the current study appeared to have similar testing experiences. Most measures of the Experience with the TA Test questionnaire yielded neutral to positive judgements for the two evaluation methods, as they also did for the HB condition. This implies that stress and awkwardness as a potential negative influence on the functionality of the testing conditions, did not play major roles in participants' experiences. Therefore, it can be said that the ecological validity of the protocols (i.e. participants being comfortable with each protocol) is ensured. Nevertheless, the HB participants did find the task of verbalising their thoughts in the retrospective phase more time-consuming than in the concurrent phase and in the other two conditions. Overall, the results suggest that while in none of the three methods was ecological validity under serious threat, usability test participants might favour the CTA or RTA method over the HB method.

Think-Aloud Methods and Usability Problems Identified

The study's results indicate that the CTA and HB methods outperformed the RTA method in terms of the quantity and quality of usability problems detected at both the individual and final problem levels. Although Ericsson and Simon [13] suggest that both concurrent and retrospective data can benefit the richness of data collected, results from the present study do not support their claim. The benefits of the HB method were not as anticipated, considering the efforts required from the participants and the evaluator. It only enabled the detection of a few more final problems, and did so at the cost of participants' experience and the evaluator's time and effort.

At the individual problem level, participants in the CTA and HB methods detected a higher number of problems than those in the RTA method, which corresponds with Peute et al.'s [32] study comparing CTA and RTA methods. It was also evident from the present study that the CTA and HB methods identified more minor problems and layout problems and elicited more problems from the verbalisation source than the RTA method. There were no significant differences found between the CTA and HB conditions in terms of the number, sources, severity levels and types of individual problems detected. The latter result conflicts with that of Følstad and Hornbaek's [14] study, which indicated that the retrospective session in the HB condition encouraged participants to identify more problems. This may be because in the aforementioned study, the researchers used interventions to specifically elicit solutions from participants, while in this study no interventions were used. At the final problem level, the CTA and HB methods detected more verbalised minor problems relating to layout problems than the RTA method. While the HB method did detect five more problems than the CTA method, these were all verbalised problems with low severity levels.

Think-Aloud Methods and Cost

No previous studies have compared the temporal cost of employing different TA methods. The findings of this study reveal that the CTA method cost substantially less than the RTA and HB methods in terms of the total time required by the evaluator to conduct testing sessions and identify usability problems. As most studies tend to compare the cost of CTA and RTA methods to other type of evaluation methods such as the heuristic evaluation method [e.g., 23, 16, 5], no comparison with previous studies can be made.

Limitations

The study participants were all drawn from one specific target group, that is, university students. While this factor has not hindered our research, it may serve to limit the application of the results to other groups who also make use of the test object, such as faculty and employees. Furthermore, the TA methods in this study were only applied to university library websites. Testing different websites with different kinds of users, such as websites aimed at elderly people, may yield results that are different from the ones presented in this thesis. It seems possible, for instance, that thinking aloud while performing tasks might present greater difficulties for elderly people than for students who have grown up with web technologies. As such, testing websites with various target groups would be very worthwhile.

CONCLUSION

This paper has discussed the results of using the traditional think-aloud methods: the concurrent think-aloud method, the retrospective think-aloud method, and the hybrid method. These three methods were compared through an evaluation of a library website, which involved four points of comparison: overall task performance, test participants' experiences, quantity and quality of usability problems discovered, and the cost of employing methods.

Overall, the findings revealed that the concurrent method can be argued to have outperformed the retrospective method and hybrid method in facilitating usability testing. It detected higher numbers of usability problems than the retrospective method, and produced output comparable to that of the hybrid method. The method received average to positive ratings from its users, and the possible reactivity associated with the concurrent think-aloud was not observed in this study, as no differences between participants' task success rates were found for this method compared to the silent condition in the retrospective test. In addition, this method required much less time on the evaluator's part than the other two methods, which required double the testing and analysis time. These findings imply a basis for preferring the concurrent method over the retrospective and hybrid methods.

ACKNOWLEDGEMENTS

We 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

1. Obead Alhadreti. 2016. *Thinking about thinking aloud: an investigation of think-aloud methods in usability testing* (Doctoral dissertation, University of East Anglia).
2. Obead Alhadreti and Pam Mayhew. 2017. To Intervene or Not to Intervene: An Investigation of Three Think-Aloud Protocols in Usability Testing. *Journal of Usability Studies*, 12(3).
3. Ali Alnashri, Obead Alhadreti, and Pam Mayhew. 2016. The Influence of Participant Personality in Usability Tests. *International Journal of Human Computer Interaction (IJHCI)*, 7(1), p.1.
4. Thamer Alshammari, Obead Alhadreti, and Pam Mayhew. 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.
5. Morten Andreasen, Villemann Henrik, Simon Schrøder, and Jan Stage. 2007. What happened to remote usability testing?: an empirical study of three methods. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 1405-1414. ACM.
6. Carol Barnum. The 'magic number 5': Is it enough for web-testing?. Information Design Journal 11, 160-170.
7. Wolmet Barendregt, Mathilde Bekker, D. G Bouwhuis, and Ester Baauw. 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.
8. John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
9. Joseph Dumas and Janice Redish. 1999. *A practical guide to usability testing*. Intellect books.
10. Maria Ebling and Bonnie John. 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.
11. K. Anders Ericsson and Herbert Simon. 1980. *Verbal reports as data.* Psychological review, 87(3), 215.
12. K. Anders Ericsson and Herbert Simon. 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge: MIT Press.
13. K. Anders Ericsson and Herbert Simon. 1993. *Protocol Analysis: Verbal Reports as Data, Revised edition*. Cambridge: MIT Press.
14. Asbjørn Følstad and Kasper Hornbæk. 2010. Work-domain knowledge in usability evaluation: Experiences with Cooperative Usability Testing. *Journal of systems and software*, 83(11), 2019-2030.
15. Wayne Gray and Marilyn Salzman. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–computer interaction*, 13(3), 203-261.
16. Layla Hasan. 2009. *Usability evaluation framework for e-commerce websites in developing countries* (Doctoral dissertation, © Layla Hasan).
17. Morten Hertzum. 2006. Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction*, 21(2), 125-146.
18. Morten Hertzum and Niels Jacobsen, N. E. 2001. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4), 421-443.
19. Kasper Hornbæk. 2010. Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, 29(1), 97-111.
20. Jonathan Lazar, Jinjuan Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.
21. Clayton Lewis and John Rieman. 1993. Task-centered user interface design. *A Practical Introduction.*
22. Gitte Lindgaard and Jarinee Chattratichart. 2007. Usability testing: what have we overlooked?. *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1415-1424). ACM.
23. Rob Martin, MA Shamari, Mohamed Seliaman, and Pam Mayhew. 2014. Remote asynchronous testing: A cost-effective alternative for website usability evaluation. *International Journal of Computer and Information Technology*, 3(1), 99-104.
24. Sharon McDonald, Helen Edwards, and Tingting Zhao. 2012. Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, 55(1), 2-19.
25. Sharon McDonald, Tingting Zhao, and Helen Edwards, H. M. (2013). Dual verbal elicitation: the complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction*, 29(10), 647-660.
26. Jakob Nielsen. 1994. *Usability engineering*. Elsevier.
27. Jakob Nielsen and Thomas Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213). ACM.

28. Jakob Nielsen .2000. Why You Only Need to Test with 5 Users. [Online] NN Group Available at https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users
29. Kenneth Ohnemus and David Biers. 1993. Retrospective versus concurrent thinking-out-loud in usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 37, No. 17, pp. 1127-1131). Sage CA: Los Angeles, CA: SAGE Publications.
30. Erica Olmsted-Hawala, Elizabeth Murphy, Sam Hawala, and Kathleen Ashenfelter. 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.
31. Erica Olmsted-Hawala and Jennifer Bergstrom. 2012. Think-aloud protocols: does age make a difference. *Proceedings of Society for Technical Communication (STC) Summit*, Chicago, IL.
32. Linda Peute, Nicolette de Keizer, and M. W. Jaspers. 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.
33. Andreas Sonderegger, Sven Schmutz, and Juergen Sauer. 2016. The influence of age in usability testing. *Applied Ergonomics*, 52, 291–300.
34. Deborah Sova, Jakob Nielsen, and NN GROUP. 2003. 234 Tips and Tricks for Recruiting Users as Participants in Usability Studies. [Online] NN Group. Available at: http://www.nngroup.com/reports/tips/recruiting/234_recruiting_tips.
35. Thomas Tullis and Bill Albert. 2008. *Measuring the user experience*. Burlington: Elsevier Inc.
36. Maaike Van den Haak, Menno de Jong, and Peter Schellens. 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.
37. Tingting Zhao, and Sharon McDonald. 2010. Keep talking: an analysis of participant utterances gathered using two concurrent think–aloud methods. In *Proceedings of the 6th Nordic Conference on Human–Computer Interaction: Extending Boundaries*. New York: ACM. pp. 581–590.
38. Tingting Zhao, Sharon McDonald, and Helen Edwards. 2012. The impact of two different think–aloud instructions in a usability test: a case of just following orders?. *Behaviour and Information Technology*, 33(2), 163–183.