Psychologie
Faculteit der Sociale Wetenschappen

The Comparison of Concurrent and Retrospective Think Aloud Methods in Unmoderated Remote Usability Testing

Name: Pelin Altuntaç
Supervisor: Pascal Haazebroek
Second reader: William L.G. Verschuur
Cognitive Psychology
Thesis Msci Applied Cognitive Psychology

Abstract

Remote usability testing has become a growing market as it is less expensive, more flexible and gives chance to conduct usability tests overseas compared to conventional usability testing. Most of the remote usability tests out in the market are unmoderated where participants have no interaction with the test conductor and they employ think aloud methods. However, advantages and disadvantages of think aloud techniques in unmoderated remote usability testing are unknown as to our knowledge comparison of think aloud techniques in remote setting has not been studied before in the literature. The present study aimed to investigate the differences between concurrent and retrospective think aloud methods in unmoderated remote usability test. Concurrent think aloud (where participants do the tasks and think aloud simultaneously) and retrospective think aloud (where participants do the tasks first and comment on the issues they encountered after completion of the test) were compared in terms of task performance, usability issues found, type of comments and participants' experiences with the think aloud method. The results showed that concurrent and retrospective think aloud are comparable in unmoderated remote usability testing. Limitations and implications of the present study were discussed.

Keywords: Concurrent think aloud, retrospective think aloud, remote usability testing.

Usability testing is a widely used method to enhance the usability of a product (Dumas & Redish, 1999). There are variety of techniques to test usability and one of the most popular one is conventional lab usability testing, where test subjects are invited to a usability lab and asked to think aloud while they are doing given tasks. However companies that are in need of usability evaluation are seeking for cheap and flexible alternatives since conventional lab methods have disadvantages such as high costs and the need for usability professionals.

Remote usability testing has emerged as a candidate for alternative to lab usability testing and as it is flexible, less expensive and gives chance to conduct usability tests overseas (Schade, 2013), it has started to be used by a lot of companies. There is a growing industry of unmoderated remote usability testing which employs concurrent think aloud such as usertesting.com and trymyiu.com. In this method, test subjects complete given tasks on their own computer at home or wherever they want and while they are doing tasks and thinking aloud simultaneously, a program records their screen, audio and/or face. Although this method seems to have advantages of remote usability testing, it might also have some potential risks due to the think aloud method. The usability literature indicated that concurrent think aloud method might lead to some issues such as prolonged reaction time, lower successful completion rate and less verbalization compared to working in silent and retrospective think aloud condition (Andersen, Hansen & Hertzum, 2009; De Jong, Schellens & Van Den Haak, 2004; Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008).

Since concurrent think aloud protocol has some issues that could alter the results and endanger validity, the retrospective think aloud protocol, where test subjects are asked to do the tasks in silent but comment on the issues that they have encountered after the completion of the test, is studied by a plenty of researchers. In

order to find out if retrospective think aloud could be equivalent to concurrent think aloud protocol without interfering the performance of subjects. There are a handful studies that makes a comparison between these two think aloud protocols (De Jong, Schellens & Van Den Haak, 2003, 2004; Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008), however all of them are in-lab techniques where experimenter sits next to the participant during the usability test. To our knowledge, there is not any research that investigated the difference of these protocols in remote settings.

This thesis project aimed to investigate the difference between concurrent and retrospective think aloud protocol in remote usability testing. The introduction is divided into four sections in order to give a detailed overview of literature related to this subject. First, it describes think aloud protocols and their potential risks, then it looks into the comparison of concurrent and retrospective protocols in conventional usability lab settings. Finally, studies pertaining to the comparison of remote and lab usability testing are provided and the aims of this thesis project are discussed.

Think Aloud Protocols

Think aloud protocols are widely used in different fields to explore people's subjective experience and thoughts through verbalization, and they have their roots in cognitive psychology. Ericsson and Simon (1993), in their classic work, stated that, "we see verbal behavior as one type of recordable behavior, which should be observed and analyzed like any other behavior" (p. 9). They discussed that verbal protocols are valid way of gathering information as long as they are collected properly. In other words, they argued that verbal protocols are not susceptible to error if participants used information available in the short-term memory and not in the long-term memory. Recently acquired information is kept in short-term memory,

whereas information in long-term memory is not directly accessible since it has to be transferred to short-term memory first.

Furthermore, they distinguished two types of think aloud protocols: concurrent and retrospective reports. In concurrent think aloud (CTA) or concurrent reports as Ericsson and Simon (1993) called, participants are asked to verbalize their thoughts while they are doing some tasks. In retrospective think aloud (RTA), participants do the tasks in silent and after completion of tasks, they are asked to verbalize their thoughts about the tasks.

Although Ericsson and Simon (1993) claimed that think aloud protocols are valid, other researchers proposed some problems that might occur due to the think aloud method. Russo, Johnson and Stephens (1989) suggested two types of invalidity that might occur due to think aloud protocols. First, think aloud protocols might interfere with the performance and prolong the reaction time. This type of invalidity is called reactivity and it is a result of change in primary process due to verbalization. Second, think aloud protocols might lead to forgetting or fabrication of some information, which is called nonveridicality.

A meta analysis of 94 studies from different fields showed that think aloud protocols do not alter the performance but it leads to prolonged reaction time (Best, Ericsson, & Fox, 2011). However, results of this study might be affected by the categorization of type of verbalization. Verbalizations were categorized as think aloud, explanatory, directed and unspecified. Researchers found that explanatory verbal protocols, where participants were asked to explain or describe while verbalizing, lead to better performance. Explanatory methods might overlap with think aloud methods, therefore it might be misleading to infer that think aloud protocols do not alter the performance to any extent and in other cases.

Concurrent vs. Retrospective Think Aloud Protocols in Usability Studies

Think aloud protocols are one of the most popular methods for usability testing. In usability studies, think aloud protocols are used to learn about users' experience with the website thoroughly. Different type of think aloud protocols are being used in the usability research but this paper only focuses on the comparison of concurrent and retrospective think aloud protocol.

In the usability literature, concurrent think aloud found to be associated with prolonged reaction time, lower successful completion rate and less verbalization (Andersen, Hansen, & Hertzum 2009; De Jong, Schellens, &Van Den Haak, 2004; Hyrskykari, Pvaska, Majaranta, Räihä, & Lehtinen, 2008).

Hertzum, Hansen, and Andersen (2007) compared concurrent think aloud and working in silent to find out if concurrent think aloud interfere with the task performance. Although they couldn't find a difference in task completion rates, participants in concurrent think aloud condition spent more time on performing the tasks than working in silent condition.

A comparative study of concurrent and retrospective think aloud protocols yielded no significant difference in terms of number and type of usability issues detected. However, participants in CTA performed lower successful than participants in RTA (De Jong, Schellens, & Van Den Haak, 2003). Another study also revealed that concurrent think aloud might result in reactivity. Eger et al. (2007) compared three think aloud protocols that are concurrent, screen cued retrospective and eye movement cued retrospective conditions. The results yielded that in concurrent think aloud condition fewer participants successfully completed the tasks compared to participants in other two conditions.

Van Den Haak et al. (2003) additionally found that in RTA, more usability issues were identified through verbalization, whereas in CTA, more usability issues were identified through observation. They explained the results by reactivity, which thinking aloud and performing the tasks simultaneously might lead to cognitive overload and consequently worse task performance and fewer verbalizations. A different study pertained to comparison of think aloud protocols also found a difference in quantity and quality of data obtained from different think aloud protocols (Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008). First of all, retrospective think aloud protocols yielded significantly more verbalizations than concurrent think aloud protocol. Second, verbalizations in concurrent think aloud method were mostly comments related to performance, whereas in retrospective conditions participants verbalized about their cognitive operations.

The usability literature also presented some contradictory findings. Van Den Haak et al. (2004) conducted a second study and compared concurrent think aloud, retrospective think aloud and constructive interaction. All methods yielded comparable results and moreover they did not find any significant difference in terms of task performance. They argued that in their first study the tasks were less difficult compared to their second study and there could be a link between task difficulty and method, even though there is no research indicating this relationship. Furthermore, Eger et al. (2007) found that screen cued retrospective condition and concurrent think aloud condition yielded equal number of usability issues, whereas eye movement cued retrospective condition produced more usability issues compared to other two conditions. They additionally found that the interaction between retrospective think aloud method and the familiarity of the website might alter the number of usability issues detected. It was found that screen cued retrospective condition produced more

usability issues when tested with a familiar search engine (Google) compared to eye movement cued retrospective condition, whereas eye movement cued retrospective condition yielded more usability issues when tested with an unfamiliar search engine (Informagnet) compared to other retrospective condition.

Lab vs. Remote Usability Testing

Usability researchers have been trying to find alternative testing methods, as conventional lab usability tests are costly, time consuming and can only be applied to limited number of test subjects due to the transportation requirements. Remote usability testing has emerged to fill this void. As it is less expensive compared to conventional methods, gives chance to conduct tests overseas and saves time for the researchers; it has become a popular method for usability testing.

The comparative studies of lab and remote usability tests showed that remote usability testing could reveal almost the same number of usability issues as in lab testing (Ames, Brush, & Davis, 2004; Bergel, Cianchette, Fleischman, McNulty, & Tullis 2002). However, the usability literature also showed that the method used by the usability practioner is important and not all remote usability testing methods are equivalent to lab usability testing. For example; Andearsen et al. (2007) compared four types of usability testing method including conventional lab testing, remote synchronous, remote unmoderated with laypeople and remote unmoderated with usability experts. They found that conventional lab testing and remote synchronous testing yielded nearly the same number of usability issues, whereas remote unmoderated methods uncovered significantly a lower number of usability issues as compared to the other methods. However, unmoderated methods in this study included written comments by participants in which they report critical incidents. To

our knowledge, there is not any research in the usability literature that employs think aloud method in unmoderated remote usability test. Therefore it is still unknown that if the unmoderated remote usability test employs concurrent think aloud is equivalent to conventional lab testing.

Current Study

This thesis project aimed to investigate the difference between concurrent think aloud and retrospective think aloud protocols in remote setting. In order to investigate the difference between think aloud methods, four research questions have been addressed.

1. Do think aloud protocols influence task completion time and task success?

   In the literature, various studies showed that CTA might result in prolonged reaction time and lower successful completion rate (Andersen, Hansen, & Hertzum, 2009; De Jong, Schellens, & Van Den Haak, 2004; Ball, Dodd, Eger, & Stevens, 2007). In the present study, participants were assigned to CTA or RTA condition and all of them were asked to do the same tasks regardless of their condition. Therefore, CTA participants were expected to spend more time to complete tasks as compared to RTA participants. Furthermore, CTA participants were expected to complete fewer tasks successfully than participants in RTA condition.

2. Is there a difference between CTA and RTA in terms of number and type of usability issues uncovered?

   A difference in the quantity of usability issues was not expected as the previous literature suggested that these two protocols reveal virtually same

number of usability issues. (De Jong, Schellens, & Van Den Haak, 2003, 2004; Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008) However, a difference in the type of usability issues was predicted as Eger et. al. (2007) found that retrospective screen cued condition uncovered more layout problems than concurrent think aloud and Van den Haak et. al.(2004) suggested that in RTA condition slightly more usability issues related to comprehensiveness was found compared to CTA condition.

3. Is there difference between CTA and RTA in terms of types of comments and number of words?

   Literature suggested that CTA participants tend to verbalize less than RTA participants. (Hyrskykari, Pvaska, Majaranta, Räihä, & Lehtinen, 2008; De Jong, Schellens, & Van Den Haak, 2003, 2004) Moreover, Hyrskykari et al. (2008) suggested that participants in the CTA condition commented largely on issues related manipulative operations (comments relating to performance) whereas participants RTA condition mostly commented on issues related to cognitive operations (Comments related to interpretations, evaluations and expectations). In the present study, it was predicted that fewer number of words would be verbalized by CTA participants compared to RTA participants. Moreover, RTA participants were expected to comment more on issues related to cognitive operations than CTA participants and CTA participants were expected comment more on issues related to manipulative operation than RTA participants.

4. How do participants evaluate think aloud protocols?

   In the usability literature, findings on participants' experience with the think aloud protocol is inconsistent. Van Den Haak et al. (2003) found that

participants rated RTA as more disturbing than CTA condition. They argued that in RTA condition, presence of experimenter during the first half of the test (when participants are asked to complete tasks in silent) might be disturbing for them. Furthermore, their research indicated that participants in RTA condition felt that they worked significantly more differently from usual than participants in CTA condition. However, Eger et al. (2007) obtained dissimilar results that participants reported that they worked significantly slower in CTA condition compared to RTA and additionally they evaluated CTA significantly more unpleasant than RTA.

As the present study employs unmoderated remote usability test, there could not be any effect of presence of experimenter on participants. However, CTA participants were expected to feel that they worked significantly more differently than their usual way of working than RTA participants, since CTA participants had to think aloud and do the tasks at the same time.

Methods

Participants

The sample was composed of 23 students from various universities in the Netherlands. The average age of participants was 22 ranging from 19 to 26. 19 of participants were female and 4 of them were male. 21 Participants were bachelor students and 2 of them were master students.

The majority of participants were recruited through Leiden University's online research participation system (SONA). Snowball sampling was also used in order to recruit more participants. All participants were randomly assigned to one of two

conditions. CTA condition consisted of 12 participants and RTA condition involved 11 participants.

Instruments

Columbia University Library Catalog (http://library.columbia.edu/) was chosen to be used as a test object for usability evaluation, as all participants were students and they were familiar with online library catalogs. None of the participants had used this website before.

All participants were given an experiment pack that included pre-questionnaire, task list, instructions paper and post questionnaire. The first item was pre-test questionnaire and it was created to gather information on demographics, participants' previous experience with library catalogs and their levels of Internet skills. The second item of the pack, task list, consisted of 3 tasks and they were as follows:

1. You need to find some publications about "child development." Please find two publications that you like and write down the names and years of the publications.
2. You need to borrow a book that was written by Sigmund Freud. Please find a book that is available today in Columbia University Libraries and write down the name and the year of the book.
3. You need to find some journal articles about "positive emotions" that were published from 1990 onwards. Please find two journal articles that you like and write down the authors and the names of the articles.

Third item in the pack was instructions paper and it was created in order to give participants information about how to use recording software. Moreover, in order to evaluate participants' experience with the particular think aloud method, a 5 point scale post questionnaire that was created by Van Den Haak et al. (2003, 2004) was used. This questionnaire originally included three sections; one more section was added for the present study. Higher scores indicated positive appraisal and sections were as follows:

1. *Appraisal of the think aloud method:* Participants were asked to evaluate their experience with the think aloud method (difficult-easy, unpleasant-pleasant, tiring-not tiring, unnatural-natural, time consuming-not time consuming)

2. *Comparison:* Participants were asked to compare the think aloud method with their usual way of working (more-less focused, more-less concentrated, more-less persevering, more-lower successful, more-less pleasant, more-less eye for mistakes, stressful-relaxed)

3. *Presence of recording equipment:* Participants were asked to rate the presence of recording equipment (unpleasant-pleasant, unnatural-natural, disturbing-not disturbing)

4. *Absence of the experimenter:* This section was not included in the original questionnaire that is used in studies of Van Den Haak et. al. (2003,2004). In this section, participants were asked to rate the absence of experimenter in the room. (Unpleasant-pleasant, feeling lost-not feeling lost, confused-not confused)

Design

This study was between subjects design. Independent variable was the think aloud condition, namely concurrent and retrospective think aloud and dependent variables were number and type of usability issues, task completion rate and time, comment type and number of words participants' experience with the particular think aloud method.

Procedure

The experiment took place in one of the labs of Leiden University Social Sciences Faculty. We have created a simulated remote usability test setting due to lack of available online facilities. During the study, participants and experimenter sat in adjacent rooms without communicating to each other in order to imitate remote unmoderated usability testing.

Firstly, participants were welcomed into lab, briefed about the procedure and asked their consent. They were given an experiment pack that consists pre-questionnaire, task list, instructions paper and post-questionnaire. Participants completed tasks and all questionnaires inside the room by their selves. After completion of the test, participants were given either 4 euros or 2 credits and they were debriefed. Detailed procedures for conditions are as follows:

Concurrent Think Aloud Condition (CTA)

Participants were briefed that they needed to think aloud while they were doing the tasks. They watched a 1-minute sample video to learn how to think aloud before they start doing tasks. During the usability test, their screen and audio input were captured.

Retrospective Think Aloud Condition (RTA)

Participants were informed that they needed to perform the tasks in silent and during the usability test their screen would be recorded. After completing the test, they were shown their own screen recording and they were asked to speak about problems they have encountered during the test. Their voice and screen were recorded while they were watching their screen recording and speaking about the problems.

Results

The data originally consisted of 26 people. However, two participants from RTA and a participant from CTA were excluded because there were not enough verbal data to analyze. 23 of the participants' (12 participants for CTA and 11 participants for RTA) recordings and questionnaires' were analyzed. A series of MANOVA was conducted to find the effect of think aloud methods on completion time, number and type of usability issues found, comment type, number of words, and participants' experiences with think aloud method. Moreover, independent samples t tests were performed to further analyze the significant results produced by MANOVA.

Completion Time and Task Success

Using Pillai's trace, there was no significant effect of think aloud condition on completion time, $V = .30$, $F(4, 18) = 1.91$, $p = .153$, although participants in CTA completed all tasks in a larger amount of time ($M = 13.87$, $SD = 8.52$) than RTA participants ($M = 11.51$, $SD = 3.59$).

Chi-square tests were performed for each task in order to find out if there is a significant association between think aloud conditions and whether or not participants completed each task successfully. The results were non-significant for task1 χ 2 (1) = 0.49, p = .48, task 2 χ 2 (1) = 0.35, p = .55 and task 3 χ 2 (1) = 0.03, p = .86. Table 1 shows the results of chi-square tests for each task.

Table 1

Crosstabulation of condition and task success

	Task 1		Task 2		Task 3	
	CS	NCS	CS	NCS	CS	NCS
CTA	11(92%)	1(8%)	9(75%)	3(25%)	7(58%)	5(42%)
RTA	9(82%)	2(18%)	7(64%)	4(36%)	6(54%)	5(46%)

*CS: completed successfully. NCS: not completed successfully.

Furthermore, an independent samples t-test was conducted in order to discover whether there is a significant association between think aloud conditions and total number tasks that were completed successfully. CTA participants completed more tasks successfully (M = 2.25, SE = 0.25) than RTA participants (M = 2, SE = 0.23). However, this difference was not significant t (21) = 0.72, p = .475.

Number and Type of Usability Issues

Recordings were analyzed in order to detect usability issues. A categorization system adopted from studies of Van Den Haak et. al. (2003, 2004) was used to classify usability issues. Categories included layout, terminology, data entry, comprehensiveness and feedback. Some examples from categories in our data are as follows:

Layout: Participant didn't see the filter options on the left side of the screen.

Terminology: Participant confused e-journal titles with articles.

Data entry: Participant didn't know how to fill in publication date in advanced search.

Comprehensiveness: Participant was not sure if the Barnard College belongs to Columbia University Libraries.

Feedback: Participant didn't understand how the search engine sorted the results.

Using Pillai's Trace, there was no significant difference between think aloud conditions in terms of total number and type of usability issues found $V = .23$, $F(5,17) = 1.01$, $p = .44$. However, in the RTA condition slightly more usability issues were found ($M = 4.45$, $SD = 3.11$) then in CTA condition ($M = 3.42$, $SD = 1.50$). Table 2 shows means and standard deviations for each of five usability issue categories.

Table 2

Type of usability issues found by participants

	CTA		RTA		
	Mean	SD	Mean	SD	Significance
Layout	2.17	1.27	2.91	2.34	n.s.
Terminology	0.92	0.79	0.64	0.51	n.s.
Data entry	0.17	0.39	0.27	0.65	n.s.
Comprehensiveness	0.17	0.39	0.45	0.69	n.s.
Feedback	0.00	0.00	0.18	0.41	n.s.

Comment Type and Number of Words

All recordings were transcribed and the number of words was calculated for each participant. In order to categorize comment types, a coding system that was adopted from Hansen, 1991 was used. This coding system consisted of three comment types that were: cognitive, visual and manipulative comments. Some example utterances in our data are as follows:

Cognitive comments:

"I guess this was the book I have to look for."

"I mixed up the e-journal titles with articles."

"That one is not correct because the author is someone else."

Visual Comments:

"I didn't see anything says library or available in library or something like that."

"I see a list of child development with only author, citation and format."

"I'm just going to look at the green ticks to see if they are available."

Manipulative Comments:

"I'm filling the search bar."

"I'm going to the home page."

"I'm going to do an advanced search."

Some utterances included more than one type of comment such as "I clicked on journal articles because I'm familiar with that." This sentence was coded as both manipulative and cognitive.

Using Pillai's Trace, there was no significant effect of condition on type of comment and number of words $V = .33$, $F\ (4,18) = 2.16$, $p = .115$. However, there was

a significant difference between conditions in visual comments. CTA participants gave more visual comments than RTA participants t (13.8) = 2.61, p = . 021. Levene's test for equality of variances was found to be violated for this analysis, F (1, 21) = 4.71, p = .042. Therefore, a t statistic not assuming homogeneity of variance was considered. Table 3 demonstrates means and standard deviations of comment types and number of words for each condition.

Table 3

Comment type and total number of words

	CTA		RTA		
	Mean	SD	Mean	SD	Significance
Cognitive	37.42	24.84	29.91	11.52	n.s.
Visual	31.83	18.40	17.09	6.36	.020
Manipulative	16.58	7.05	13.36	9.00	n.s.
Total no of words	639.00	249.12	540.00	234.52	n.s.

Participant Experience

A series of MANOVA were performed to analyze participants' experience with think aloud methods. There was not a significant effect of think aloud conditions on appraisal of the method V = .16, F (5,17) = .65, p = .664.

There was also not an overall significant effect of condition when participants compared the usability method to their usual way of working V = .50, F (8,13) = 1.65, p = .203. However, RTA participants felt that they worked more successfully than their usual way of working compared to CTA participants. This effect was significant t (21) = -2.11, p = .048.

Furthermore, the effect of think aloud condition were non-significant for presence of recording equipment $V = .00$, $F\ (3,19) = .02$, $p = .995$ and absence of experimenter $V = .03$, $F\ (3,19) = .18$, $p = .912$.

Discussion

The present study aimed to find the differences between concurrent and retrospective think aloud protocols in unmoderated remote usability testing. Results showed that these methods are comparable in terms of task performance, usability issues found, quantity and quality of the comments and participants' subjective experience with the think aloud method.

There are four major findings of this study. First of all, unlike predicted, CTA didn't cause reactivity. There are contradictory findings on reactivity of CTA in the literature. Several studies found that CTA led to prolonged reaction time (Hertzum, Hansen, & Andersen, 2009) and lower successful task completion rate (De Jong, Schellens, & Van Den Haak, 2003; Ball, Dodd, Eger, & Stevens, 2007). However, there are also other studies which are consistent with our findings. In a study of Hertzum et al. (2009), CTA didn't lead to lower successful completion rate. Furthermore, some studies didn't find a significant difference between CTA and RTA in terms of completion time (De Jong, Schellens, & Van Den Haak, 2003, 2004).

Secondly, results yielded that RTA and CTA are comparable in terms of number and type of usability issues uncovered. The previous studies pertained to comparison of CTA and RTA also are consistent with the present study. They found that both CTA and RTA revealed virtually same number of usability issues. (De Jong, Schellens, & Van Den Haak, 2003, 2004; Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008) However, unlike predicted, there was no significant difference between

think aloud methods in terms of type of usability issues. The results of former studies showed differences in layout and comprehensive usability issues between two think aloud methods. (De Jong, Schellens, & Van Den Haak, 2004; Eger, Ball, Stevens, & Dodd, 2007) In the present study, the only difference that was found regarded the type of usability issues was feedback that RTA participants found two feedback issues whereas CTA participants did not encounter any feedback issues. However, this finding was not significant.

Thirdly, the most surprising result of the study was that CTA condition produced more verbal data than RTA condition, even though the difference was not significant. This finding was inconsistent with previous literature, since they found RTA condition produced significanlty more verbal data than CTA condition (Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008; De Jong, Schellens, & Van Den Haak, 2003, 2004). For higher verbalization in CTA condition, two possible explanations are proposed: (1) in present study, CTA participants were shown a 1-minute sample video to learn how to think aloud whereas RTA participants were not which might lead to fewer verbalizations in RTA condition as they didn't know how to think aloud, (2) in one study which proved that RTA led to more verbalization than CTA, (Hyrskykari, Lehtinen, Majaranta, Pvaska & Räihä, 2008) participants were also shown their gaze paths in RTA condition which might elicited more verbalizations.

Another unexpected finding was the difference in type of comments. CTA condition elicited more visual comments than RTA condition and unlike predicted, there wasn't any significant difference between think aloud condition in cognitive and manipulative comments. This finding was inconsistent with previous studies. Hyrskykari et al. (2008) found that there was not any significant difference between

think aloud conditions regarding visual comments and RTA conditon produced more cognitive comments whereas CTA condition produced more manipulative comments. Two possible explanation are proposed for this difference: (1) in present study, it was observed that in CTA condition participants read the contents of the website (such as titles, button names, section names etc.) while they were doing the tasks and these utterances were coded as visual comments. (2) In the study of Hyrskari et al. (2008), RTA participants were also shown their gaze paths which might elicited more cognitive comments.

Lastly, RTA participants significantly felt more successful than CTA participants. One possible explanation for this finding is that RTA participants watched their recordings and had the chance to see the things that they did correctly which might make them certain about their success. There are contradictory findings on participants' experience in the literature, however one finding was consistent with the present study which they found RTA participant felt that they worked significantly more differently than their usual way of working (De Jong, Schellens, & Van Den Haak, 2003).

Limitations

There are several limitations of this study. Firstly, due to lack of available online facilities a simulated remote lab condition was designed. Even though participants and experimenter were not in the same room during the experiment, participants were not exposed to any interruptions. However, in actual remote setting where users are free to choose a place that they complete usability test, they might be interrupted and literature suggests that in the presence of interruptions, think aloud protocols tend to be more reactive (Hertzum & Holmegaard, 2013).

Secondly, the number of participants might not be enough to investigate the difference between two think aloud conditions. Although Nielsen (1994) asserted that 6 or 7 participants would be enough to detect %75 of usability issues, this claim might not be valid to uncover difference between usability evaluation methods. Caulton (2001) claimed that homogeneity of variances would be violeted in usability studies if the number of participants are not enough. In the present study, homogeneity of variances were also violated for several analysis which might led to inaccurate results.

Lastly, participants were asked to speak in English which was not their first language. Even though none of the participants were native English speaker in order to prevent biases, using second language instead of native language might have altered the results.

Implications

This study revealed that CTA and RTA might be equivalent in remote usability testing, as they uncover almost same number of usability issues and the task performance are not affected by the think aloud method. Therefore, CTA might be more efficient to use, as it requires less time and no additional software. However, future studies should consider doing different type of RTA such as eye movement cued RTA as previous studies revealed that there might be differences between different types of RTA and CTA (Ball, Dodd, Eger, & Stevens, 2007). Moreover, future studies should recruit more participants to prevent violation of homogeneity of variances.

Conclusion

There are a lot of companies that employ think aloud methods in remote settings. However, their method of usability evaluation lacks validity since there is not enough academic study that investigates methods currently used in remote usability testing. More studies should explore disadvantages and advantages of these methods in order to preserve validity.

References

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What Happened to Remote Usability Testing? An Empirical Study of Three Methods . *CHI 2007*, (pp. 1405-1414). San Jose, California.

Brush, A. B., Ames, M., & Davis, J. (2004). A Comparison of Synchronous Remote and Local Usability Studies for an Expert Interface . *CHI 2004*, (pp. 1179 1182). Vienna, Austria.

Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing . *Behaviour & Information Technology , 20* (1), 1-7.

Dumas, J. S., & Redish, J. C. (1999). *A Practical Guide to Usability Testing.* Exeter, England: Intellect.

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing Retrospective Verbal Reports in Usability Testing Through Eye-Movement Replay. *Proceedings of HCI 2007* (pp. 129-137). British Computer Society.

Ericsson, A. K., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data* (Revised Edition ed.). Cambridge, Massachusetts, United States of America: The MIT Press.

Fox, M. C., Ericsson, A. K., & 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), 316-344.

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica , 76*, 31-49.

Hertzum, M., & Holmegaard, K. D. (2013). Thinking Aloud in the Presence of Interruptions and Time Constraints . *Intl. Journal of Human–Computer Interaction , 29*, 351-264.

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

Hyrskykari, A., Pvaska, S., Majaranta, P., Räihä, K.-J., & Lehtinen, M. (2008). Gaze Path Stimulation in Retrospective Think-Aloud . *Journal of Eye Movement Research , 2* (4), 1-18.

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *Int. J. Human-Computer Studies , 41*, 385-397.

Russo, E. J., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition , 17* (6), 759-769.

Schade, A. (2013, October 12). *Remote Usability Tests: Moderate and Unmoderated*. Retrieved June 1, 2015, from NNGroup: http://www.nngroup.com/articles/remote-usability-tests/

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. (2002). An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. *Usability Professionals Conference.*

Van Den Haak, M. J., De Jong, M. D., & Schellens, P. J. (2004). Employing Think Aloud Protocols and Contructive Interaction to Test The Usability of Online Library Catalagues: A Methodogical Comparison. *Interacting with Computers ,16*, 1153-1170.

Van Den Haak, M., De Jong, M. D., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue . *Behaviour & Information Techology , 22* (5), 339-351.