View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Sunderland University Institutional Repository

Willis, Leanne and McDonald, Sharon (2016) Retrospective protocols in usability testing: a comparison of Post-session RTA versus Post-task RTA reports. Behaviour & Information Technology, 35 (8). pp. 628-643. ISSN 0144-929X

Downloaded from: http://sure.sunderland.ac.uk/6284/

Usage guidelines

Please refer to the usage guidelines at http://sure.sunderland.ac.uk/policies.html or alternatively contact sure@sunderland.ac.uk.

Taylor & Francis
Taylor & Francis Group

Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: http://www.tandfonline.com/loi/tbit20

Retrospective protocols in usability testing: a comparison of Post-session RTA versus Post-task RTA reports

Leanne M. Willis & Sharon McDonald

To cite this article: Leanne M. Willis & Sharon McDonald (2016): Retrospective protocols in usability testing: a comparison of Post-session RTA versus Post-task RTA reports, Behaviour & Information Technology, DOI: 10.1080/0144929X.2016.1175506

To link to this article: http://dx.doi.org/10.1080/0144929X.2016.1175506

Published online: 12 May 2016.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tbit20

Download by: [University of Sunderland] **Date:** 12 May 2016, At: 08:33

BEHAVIOUR & INFORMATION TECHNOLOGY, 2016
http://dx.doi.org/10.1080/0144929X.2016.1175506

Taylor & Francis
Taylor & Francis Group

Retrospective protocols in usability testing: a comparison of Post-session RTA versus Post-task RTA reports

Leanne M. Willis and Sharon McDonald

Department of Computing, Engineering and Technology, Faculty of Applied Sciences, David Goldman Informatics Centre, University of Sunderland, Sunderland, UK

ABSTRACT

We present the results of a study that compared two placements of the Retrospective Think-aloud (RTA): A Post-session RTA where the think-aloud occurs after all tasks are complete, and a Post-task RTA where the think-aloud is elicited after each task. Data from task performance and verbal measures were collected from 24 participants. The results suggest that in terms of task performance, participants in the Post-session RTA condition performed tasks faster, with fewer errors and fewer clicks than in the Post-task RTA condition. In terms of utterances, participants in the Post-task RTA condition produced significantly more utterances that explained actions, expectations and procedural descriptions than in the Post-session RTA condition.

ARTICLE HISTORY
Received 3 September 2015
Accepted 1 April 2016

KEYWORDS
Usability testing; think-aloud studies; Retrospective Think-aloud

Downloaded by [University of Sunderland] at 08:33 12 May 2016

1. Introduction

In recent years, the elicitation procedures that underpin the use of think-aloud protocols have received much scrutiny. The concurrent think-aloud has been the primary focus of this interest (Hertzum, Hansen, and Andersen 2009; Olmsted-Hawala et al. 2010; McDonald and Petrie 2013; Zhao, McDonald, and Edwards 2014; McDonald, Zhao, and Edwards 2015). The Retrospective Think-aloud (RTA), however, has received less attention (Guan et al. 2006; Eger et al. 2007; Elling, Lentz, and de Jong 2011). Elicitation procedures for the RTA frequently involve users providing their verbal protocols about tasks in a single block at the end of the test; usually cued by a muted video replay of the session (Bowers and Snyder 1990; Page and Rahimi 1995; van den Haak, de Jong, and Schellens 2003). However, the greater the interval between task completion and retrospective recall, the more likely it is that the accuracy of the RTA will suffer: users may simply forget the reasons for their behaviours, they may generalise across tasks or indeed rationalise their behaviours (Ericsson and Simon 1984, 1993; Taylor and Dionne 2000). Generalisations and rationalisations are a particular concern for usability testing; generalisations are unlikely to help evaluators to diagnose specific issues and rationalisations may threaten accurate problem diagnosis.

In this paper we investigate the impact of two different placements of RTA: a Post-session RTA which takes place after all tasks are complete, and a Post-task RTA which takes place after each individual task is complete, on the nature of the think-aloud protocols produced, and task performance. Before we describe our study and report our findings, we briefly review the literature pertaining to the use of the RTA within usability testing.

1.1. The RTA: its place in usability testing

In their influential work on protocol analysis, Ericsson and Simon (1984, 1993) suggest that both concurrent and retrospective protocols should be collected. The concurrent think-aloud is to understand task-based cognitive processes and the RTA is to gain explanatory insights. However, within usability testing research, the concurrent think-aloud and RTA have emerged as separate techniques. The concurrent technique is reported as being used more frequently by practitioners (McDonald, Edwards, and Zhao 2012), despite evidence that the retrospective technique generates the type of explanations and reflections that practitioners find to be particularly useful (Bowers and Snyder 1990; Ohnemus and Biers 1993; van den Haak, de Jong, and Schellens 2003). The speed and simplicity of concurrent elicitation coupled with the immediacy of feedback means that the concurrent technique is well suited to situations, such as usability testing, where results are often needed within a short timeframe. Given the practical benefits of the concurrent technique, and that the RTA does increase the length of tests sessions, why might practitioners consider the retrospective approach? We suggest the answer to this question is twofold. First, questions have been raised about

CONTACT Leanne M. Willis leanne.willis@research.sunderland.ac.uk
© 2016 Informa UK Limited, trading as Taylor & Francis Group

Downloaded by [University of Sunderland] at 08:33 12 May 2016

the validity of concurrent reports. Within usability testing, the focus of attention rests on divergent practice between Ericsson and Simon's classic concurrent procedures and how practitioners gather concurrent protocols during commercial usability testing (Boren and Ramey 2000; Nørgaard and Hornbæk 2006; Shi 2008; McDonald, Zhao, and Edwards 2015).

Ericsson and Simon (1984, 1993) argue that the validity of concurrent protocols is dependent upon the elicitation procedures used. When elicitation procedures require users to verbalise thoughts that go beyond the moment-by-moment cognitive processes involved in task execution, for example, by asking users to reflect upon the reasons for actions, there is a risk that these higher-order thought processes may lead to an artificial change in task performance. This change in performance is referred to as reactivity and it may render the concurrent protocol invalid (van den Haak, de Jong, and Schellens 2003; Hertzum, Hansen, and Andersen 2009; Olmsted-Hawala et al. 2010; Fox, Ericsson, and Best 2011; McDonald and Petrie 2013). By contrast, because retrospective protocols are elicited after tasks are complete, the RTA facilitates the elicitation of explanations and reflections without influencing what users do during task performance. Therefore, it should sidestep some of the issues associated with its more popular, but troublesome, sibling.

Second, studies that have examined the content of the verbal data produced during the concurrent technique suggest that participants provide the type of verbalisations associated with reactivity (explanations and reflections) even when classic administration procedures are used. That is, the think-aloud procedures focused only on eliciting task-based cognitive processes rather than higher-order cognitive processes (Zhao and McDonald 2010; Hertzum, Borlund, and Kristoffersen 2015). Zhao and McDonald (2010) suggest that the context of usability testing may override the explicit instructions to think-aloud. Within a usability study, users are aware that the product is the focus of the evaluation and therefore they might think that their opinions, reflections and recommendations are required even when they are not directly solicited. Consequently, it may be that the production of reflections and explanations during the concurrent think-aloud are difficult to avoid. Indeed, a meta-analysis of the results of think-aloud studies from cognitive psychology concluded that the classic concurrent think-aloud was not reactive beyond extended task completion times (Fox, Ericsson, and Best 2011). However, there is evidence from studies within usability testing to suggest that, even when classic administration procedures are used, reactivity may result (van den Haak, de Jong, and Schellens 2003; Hertzum and Holmegaard 2013). Such findings may, indeed, reflect the contextual differences between usability testing and psychology experiments. However, further research is needed to isolate those conditions in which the classic method may be reactive in usability testing research.

Studies investigating the contribution of retrospective reports to usability testing have done so, in the main, using between-subjects comparison of the retrospective and concurrent think-aloud (see e.g. Bowers and Snyder 1990; Ohnemus and Biers 1993; van den Haak, de Jong, and Schellens 2003) and between the retrospective, concurrent and team-based approaches such as constructive interaction (van den Haak, de Jong, and Schellens 2004, 2007, 2009). These studies suggest that when compared with the concurrent think-aloud, the retrospective technique yields utterances that have more value for usability analysis (Bowers and Snyder 1990; Ohnemus and Biers 1993) and more verbalised usability problems (van den Haak, de Jong, and Schellens 2003, 2004).

However, the findings with respect to the differences between the concurrent and RTA are not always consistent. For example, the increased detection of problems from verbal data in the RTA reported by van den Haak, de Jong, and Schellens (2003, 2004) was not confirmed in two follow-up studies (2007 and 2009). Both studies did report that, in terms of the number of individual problems detected per method, the RTA and constructive interaction out-performed the concurrent technique.

However, van den Haak, de Jong, and Schellens (2007) report that the RTA gave rise to more observable usability problems than constructive interaction (no differences were found between the retrospective and concurrent methods, or indeed concurrent and constructive interaction). Participants in the retrospective method were also less successful in terms of task completion than constructive interaction. van den Haak, de Jong, and Schellens (2007) attribute these differences, and the reduction in the number of verbalised problems in the retrospective condition, to the characteristics of the products used during the different tests. Where the site architecture requires users to spend significant periods of time engaged in reading activities the RTA might be less useful for evaluative purposes. On a practical level, the video cue of participants' test session included fewer retrieval cues when reviewing reading activities, and the poorer task performance suggests that the retrospective condition may have skim read text that was pivotal in subsequent navigation decisions and, as a consequence, experienced more problems.

Studies investigating the combined use of the concurrent and RTA within the same test suggest that retrospective reports can generate insights into the reasons

Downloaded by [University of Sunderland] at 08:33 12 May 2016

behind encountered difficulties and decisions made during task performance. Page and Rahimi (1995) found that the concurrent think-aloud generated significantly more procedural statements whilst the retrospective generated more explanatory statements. The retrospective phase also produced more statements relating to errors of strategy made by users. McDonald, Zhao, and Edwards (2013) report that the RTA helped to shed light on issues identified by the concurrent technique by reinforcing the impact of an issue; explaining the causes of encountered difficulties; and providing contextual information about the impact of encountered difficulties and usability issues that were not verbalised during the concurrent session. However, McDonald, Zhao, and Edwards (2013) also report evidence of a small number of undesirable retrospective utterance types including: hypothesising, rationalising and forgetting. In the next section we briefly discuss other possible validity concerns with the retrospective method and their relationship to elicitation procedures.

1.2. The validity of retrospective reports

The primary validity concern for the RTA is that it relies upon the user's memory of their task-solving process and memories are not necessarily veridical. Indeed, concerns about the validity of retrospective reports relate to the specificity and validity of the information provided. A number of validity issues with retrospective reports have been identified (Taylor and Dionne 2000; Ericsson and Simon 1984, 1993). These are discussed below:

(1) Generalisation: Retrospective reports may refer to general episodes rather than task-specific episodes. Distinctive memory traces are easier to retrieve than memories that bear a close resemblance to one another. Therefore, Ericsson and Simon (1984, 1993) suggest that generalisations are more likely to occur when participants are asked to solve a number of similar tasks in close succession. This situation may be exacerbated in usability testing contexts where participants are asked often to complete numerous tasks with the same product.
(2) Invention: Participants may invent thoughts they did not have during the test (van den Haak, de Jong, and Schellens 2003; Eger et al. 2007). This might, for example, include reasons for individual actions or strategic approaches to task completion.
(3) Rationalisation: Participants may attempt to explain or justify their behaviour with logical, plausible reasons that may not necessarily reflect the truth. The use of video to cue retrieval might add to this problem in that participants may respond to elements of the visual stimuli rather than confining their report to their memory of task performance (Leow 2002; Cotton and Gresty 2006).

Retrospective reports may also suffer from some of the issues that affect concurrent reports: filtering and editing. Participants may be selective about what information they report in their think-aloud and filter the information they provide. For example, participants may choose not to disclose certain pieces of information if they think the experimenter has an interest in the product being used in the test (Eger et al. 2007).

In an effort to increase the accuracy of the RTA, memory cues such as a video replay of the test session are generally (but not always) used as a mechanism to ground the protocol to the users' actual performance (Bowers and Snyder 1990; Page and Rahimi 1995; van den Haak, de Jong, and Schellens 2003). Some researchers have also investigated supplementing this with traces of the users' eye movements, although no differences were found in problem detection rates between video replay and eye-cued RTA (Eger et al. 2007; Elling, Lentz, and de Jong 2011).

Despite justified concerns over the validly of retrospective reports, there is evidence to suggest that they are accurate. Guan et al. (2006) examined the congruence of retrospective reports with users' eye movements collected during the completion of four tasks. Guan et al. (2006) found the verbalisations to be an accurate true reflection of what participants did during task performance with only 3% of verbal reports being inaccurate. Approximately half of the verbalisations were about procedures with around one-third of utterances relaying useful explanatory data. However, the tasks used were similar to the type of tasks used in psychological investigations of verbal protocols rather than the type of tasks used in usability testing. McDonald, Zhao, and Edwards (2013) report that despite not cuing recall with a replay of the test session, inaccurate recollections (instances of forgetting) accounted for only 3% of the utterances made in their retrospective condition. They suggest that their varied task set may have helped to mitigate this problem, helping participants to distinguish between tasks.

1.3. Elicitation procedures and the RTA

As with the concurrent technique, the elicitation procedures used during a RTA are likely to be a key factor in determining its validity (Taylor and Dionne 2000). Researchers considering the use of the retrospective technique will face the choice of a number of elicitation options including: the use of retrieval cues; instruction types; evaluator probes and the placement of think-

Downloaded by [University of Sunderland] at 08:33 12 May 2016

aloud. Ericsson and Simon (1993) suggest that, in an ideal world, participants should perform the RTA immediately after tasks have been completed; thereby ensuring that the required information should still be in short-term memory. A significant delay between task performance and verbal reporting is likely to erode the accuracy of a participant's memory trace.

Most investigations of the RTA have confined the think-aloud to one single session at the end of the test rather than reporting after each separate task. However, reporting after each task may be beneficial, and task-by-task reporting is used in approaches such as Co-operative Usability Testing (CUT) (Frøkjær and Hornbæk 2005; Følstad and Hornbæk 2010). Følstad and Hornbæk (2010) suggest that interpretation sessions after each task should provide instant access to a test user's interpretation of the system and that users would be less likely to try and rationalise their behaviour. Følstad and Hornbæk (2010) extended the CUT method to include an interpretation session after each individual task was completed rather than after the final task. The interpretation session was structured as a task walk-through rather than a video-cued, task-based discussion as in the original CUT method (Frøkjær and Hornbæk 2005). Their findings suggest that the interpretation phases generated new usability problems and provided additional insights about issues that had already been observed.

Although Følstad and Hornbæk's (2010) extension to the CUT method was an evaluator-led interpretation session rather than a RTA, their findings suggest that for approaches which seek to understand the user experience task-by-task think-aloud may bear dividends. Indeed, during a usability test, participants are often required to complete lengthy task sets and therefore one might expect their memory for the detail of specific tasks to erode. Moreover, because users are performing tasks with the same test product, continued exposure is likely to affect the distinctiveness of individual tasks; thereby further increasing the possibility of unhelpful utterance types such as generalisations and rationalisations.

1.4. The present study

The study presented here examined the impact of two different placements of the RTA: a Post-session RTA in which protocols are elicited after all tasks are complete and a Post-task RTA in which protocols are elicited after each individual task. We investigate the following hypothesis:

As tasks are completed in silence for both the Post-session and Post-task RTA we expect no differences across the task performance measures.

> H1: Think-aloud placement will have no effect on task performance measures.

The proximity between action and recall in the Post-task RTA condition should mean that users' recollections of the things that caused them difficulty or delight should still be in short-term memory. We therefore might expect the protocols within the Post-task RTA condition to contain more detailed procedural descriptions, explanations and utterances that convey insights into the users' experience.

> H2: The Post-task RTA will lead to an increase in the number of utterances made about users' task-solving behaviours over the Post-session RTA.

> H3: The Post-task RTA will lead to an increase in the number of utterances made about the user experience (positive and negative), user expectations and explanations of behaviour over the Post-session RTA.

2. Methodology

In this section we describe the design and the test procedure that we followed in our study. Permission to run the study was sought and granted from our University Research Ethics Committee.

2.1. Participants

Twenty-four volunteers participated in the study: 12 males and 12 females. Their ages ranged from 18 to 64 years, with a mean age of 33 years. Participants were drawn from staff and students at a university in the North East of England. All participants were representative users of the test products as determined by their responses to a user profile questionnaire. All of the participants reported that they were frequent users of the Internet, with 92% of the participants stating they used the Internet several times a day. Participants received no incentives to participate in the study.

2.2. Materials and tasks

Two museum websites were used in this study: website A the natural history museum (www.nhm.ac.uk) and website B the science museum (www.sciencemuseum.org.uk). These sites were selected because they have a broad user base, which helped to facilitate the recruitment of representative users. Moreover, because these sites contained the same types of elements (e.g. visitor information and an online repository about exhibits and subject matter) we were able to match tasks in terms of both focus and difficulty.

Downloaded by [University of Sunderland] at 08:33 12 May 2016

The first author developed six tasks for each website; all tasks were piloted before testing to ensure their wording was clear and free from bias. To mitigate the effects of learning in relation to the test products, each task was focused on a different branch of the sites' information hierarchy. The tasks used related to planning a visit or finding information about museum exhibits. We matched tasks in terms of both difficulty and focus between the websites. In terms of difficulty, tasks were matched by ensuring their solutions were at the same level of depth within each site. In terms of focus we matched tasks in terms of the type of information users were asked to find. Each task had one correct answer. For example:

> Website A
>
> When is the next Dino Snores event taking place?
>
> Website B
>
> When is the next science Night taking place??

2.3. Study design

The study used a repeated measures design with an independent variable of think-aloud placement. The independent variable had two levels: A Post-session RTA, where participants were asked to think-aloud at the end of all tasks, and a Post-task RTA, where participants were asked to think-aloud after each task. Half of the participants started with the Post-session RTA condition, while the other half started with the Post-task RTA condition. Within each placement of the think-aloud, half of the participants started with Site A, while the other half started with Site B (see Table 1); this measure mitigated the risk of order effects. We randomly assigned participants to one of the four testing groups before their arrival at the laboratory.

Regardless of condition, participants were told that they would be required to provide an RTA before they began their tasks. This is counter to some studies where participants were told of the need to think-aloud only after the tasks are complete (see e.g. Guan et al. 2006; Eger et al. 2007; Elling, Lentz, and de Jong 2011). If we had not forewarned participants, in both conditions, that an RTA would be requested then the two conditions would have been imbalanced. During the Post-session RTA, participants would have completed all tasks oblivious to the need to think-aloud afterwards. However, in the Post-task RTA participants would have completed only the first task under these conditions, as following the first Post-task RTA participants would have come to expect the need to verbalise after each task. Therefore, we believed that telling participants about the coming RTA before tasks were attempted was the only way to compare the two think-aloud placements.

Table 1. Study design.

	INSTRUCTIONS	Post-session RTA	BREAK	Post-task RTA	Ratings
P1–6		Site A		Site B	and
P7–12		Site B		Site A	Interview
		Post-task RTA		Post-session RTA	
P13–18		Site A		Site B	
P19–24		Site B		Site A	

2.4. Study procedure

The test sessions took place in our usability laboratory and were facilitated by the first author. Each session was conducted on a one-to-one basis and lasted around 1 hour 15 minutes including instructions and debriefing.

Following the completion of all necessary consent forms, the test facilitator explained the purpose of the study. The participants were told they would be helping to evaluate two different websites. At this point participants were not told the purpose of the study was to investigate the placement of the think-aloud within the test session; this was, however, explained to them at the end of the second evaluation during the debriefing session. The facilitator took time to make sure participants were comfortable and at ease before starting the tasks and made sure to highlight that the study was an evaluation of the products and not the user. Within each condition we confined the information we communicated to participants about only that condition; we did not tell them what to expect in the second evaluation.

In both conditions, participants were told that they would be completing six tasks with the test website. They were also told that they would be asked to provide a video-cued think-aloud. We asked participants to complete the tasks without help, but if they felt during the completion of a task that they would not persist in real life then they could abandon that task. Once the facilitator noted that the participant understood what was required of them, she handed over the first task to the participant and the test began. Tasks were handed one-by-one to participants as they progressed through each evaluation. The test facilitator sat in the room with the participant a little way behind them and to the right-hand side.

In the Post-session RTA condition, participants provided their think-aloud after completing the last of their six tasks using the following instructions:

> I am now going to show you the test video of your session. As the video plays I would like you to recall the

thoughts you had when you completed each task and say them out loud. If you are silent for any length of time I will remind you to keep talking. If you have any questions please ask them now; if not, you may begin.

In the Post-task RTA condition, participants provided their think-aloud after each task. The same think-aloud instruction was used as in the Post-session RTA with a slight wording modification to reflect the focus on a single task: 'I would like you to recall the thoughts you had when you completed the task'. For both conditions, no evaluator probes were used during task completion or during think-aloud elicitation.

After thinking-aloud in each condition we asked participants to complete a short (three item) Likert scale consisting of the following questions: the content of my think-aloud was accurate; I relied heavily on the video replay while thinking-aloud and I remembered all of the tasks. The scale used to rate these questions was a 1 to 5 scale where 1 was Strongly Disagree and 5 was Strongly Agree. After the evaluation was complete participants helped us to understand their experiences with the think-aloud approaches through a brief semi-structured interview. Finally, the test facilitator debriefed on the purpose of the study and thanked participants for their time. Both sessions were conducted on the same day separated by a half-hour rest period.

2.5. Dependent measures

Verbal data: the number of utterances made and the nature or type of utterance produced.

Task performance: time on task, task success, the number of mouse clicks (these included within and between page clicks), and the number of errors made. We identified two types of error: slips and divergences. We define a slip as an accidental error, a mistake, that is recognised by participants during task performance and that was immediately rectified (within 15 seconds). For example, a participant might accidently select the wrong item from a list but immediately correct themselves. A divergence was counted when participants made an incorrect link selection that was not accidental and that was not immediately corrected.

The first author undertook the data coding for both types of error. In coding these errors, a simple checklist (see Table 2) was devised to provide structure to the coding process. This checklist included observational and verbal indicators. Before any coding was started, all possible routes to the answer were recorded.

Slips were coded as follows: the first author watched the test videos with the possible task solution routes in front of her along with the checklist. Each time an item on the coding scheme was detected, 15 seconds was counted using a stopwatch. If a participant corrected himself or herself within the 15-second time limit then this was counted as a slip. A frequency count was recorded for the number of slips made (and indicator type) by each participant in each condition.

Table 2. Checklist used to guide the identification of slips and divergences.

Indicator	Definition
Indication types based on observed behaviour	
Wrong link	Participant clicks on the wrong link
Missed link	Participant misses a step in the navigation process
Repeated action[a]	Participant clicks the same link they have already tried
Indication types based on verbalised behaviour	
Recognition[a]	Participant realises they have made a mistake by verbalising, for example: 'I didn't mean to do that' or 'that's not right'
Random action[b]	Participant verbalises that they are now performing a random action
Wrong understanding[b]	The participant verbalises an incorrect understanding of site features, for example, a link, text, terminology

[a]Only applies to slips.
[b]Only applies to divergences.

Divergences were coded as follows: the first author watched the test videos with the possible task solution routes in front of her along with the checklist.

Each time an item on the coding scheme was detected, 15 seconds were counted using a stopwatch. If a participant did not correct him- or herself within the 15-second time limit then this was counted as a divergence. When counting divergences, each subsequent click in the route was counted as a separate divergence. A frequency count was recorded for the number of divergences (and indicator type) made by each participant in each condition.

The first author independently coded all of the data. Following a period of one month she repeated the coding again on a sub-set of the data from 36 tasks. Cohen's Kappa was calculated as a measure of reliability and a good level of agreement was found, slips ranged from 0.66 to 0.88 with an average of 0.76, and divergences ranged from 0.61 to 0.81 with an average of 0.70.

2.6. Qualitative analysis process

We present two types of qualitative data in this paper. The first is an analysis of participants' verbal data. The second is an analysis of participants' interview data. In the following sections, we describe how we coded both the verbal and interview data.

2.6.1. Verbal data

The first author transcribed all of the test sessions. Transcription was conducted approximately three weeks following data collection and analysis commenced a further five weeks later. The transcripts were segmented into

Downloaded by [University of Sunderland] at 08:33 12 May 2016

Downloaded by [University of Sunderland] at 08:33 12 May 2016

individual utterances and to each utterance an interpretative code was attached. Utterances could vary in length but each focused on a single topic. We used Context Appreciative Coding (Yang 2003); this approach involves simultaneous segmentation and coding. The surrounding utterances were examined and where necessary test videos were revisited. We believe that checking the context in which an utterance occurred helped with coding accuracy.

The coding scheme was inspired by that used by McDonald, Zhao, and Edwards (2013). However, we were open to the possibility that new codes could emerge. There were several differences between our coding scheme and that used by McDonald, Zhao, and Edwards (2013); we believe these differences emerged because our test utilised a video replay of the test session whereas McDonald, Zhao, and Edwards's (2013) scheme was based on free recall. We kept the category of 'Procedural Description' but felt it was necessary to extend this further by having separate categories for 'Scanning' and 'Scrolling'. Three new categories, Text summary, Video Cue and Technical Problems, were added to the scheme. Table 3 presents the final coding scheme with examples of each category.

To foster consistency and understanding of the coding scheme both authors coded and segmented the first transcript independently and then discussed their coding and resolved any differences. The first author then segmented the remaining transcripts.

The first author coded the remaining transcripts independently and the second author independently coded one in every 4 sessions (12 in total). This provided a measure of coding reliability using Cohen's Kappa (0.81). This demonstrated good coding reliability.

The 35 remaining transcripts were crosschecked by the second author and disagreements were discussed. The second author was given all of the segmented and coded transcripts and a copy of the coding scheme. She was unaware from which session each transcript came. In total, 96 out of 2693 (4%) utterance codes were changed following this process. As the remaining transcripts were crosschecked, rather than being independently coded, they were not included in the Kappa analysis presented above.

2.6.2. Interview data

The first author transcribed all of the interview data into individual files. The authors used open coding to analyse the individual transcripts. The authors worked together on the first two transcripts, in two separate analysis sessions. Before meeting to begin coding, each author independently read through the transcripts several times to

Table 3. Coding scheme for utterance data.

Category name	Definition	Example
Procedural description[a]	Read out text, links; describe what they were doing, trying to do or did	'So firstly went back to the homepage, start at the beginning, just having a look around really'. P1
Scanning	Describing visual behaviour i.e. scanning, glancing, looking	'I spent most of my time re-reading over it, over and over again, skimming over it'. P2
Scrolling	Describing use of the scroll bar	'Then scrolled down to look at some more information on the homepage'.
Action explanation[a]	Explain the reason(s) for executing or going to execute certain actions	'I clicked on this one because it says Archives and Collections and it had Objects so I thought this was the best one'. P12
Text summary	Summarising information they have read	'But it was just talking about things relating to climate change'. P19
Expectation[a]	Express what is/was going to happen, including anything counter to expectations	'I clicked on Announcements because I thought there might be something in there'. P24
Positive user experience	Expression of positive feelings and experience caused by the site.	'Clicked on Country which was quite easy to find'. P13
Negative user experience	Expression of negative feelings and experience caused by the site.	'But it wasn't very practical because obviously there is a lot of scrolling needed'. P10
Usability issue	Description of an experienced issue with respect to dialogue functionality, layout or navigation	'I didn't know how this filter worked, normally when you select something it should come up but whatever you selected didn't come up'. P14
Recommendation[a]	Give recommendations on how to improve the interface	'Like an event list with a brief description and where they actually are as well'. P21
Performance assessment[a]	Difficulty or ease of solving a task; time on task; whether or not the correct answer was found.	'So it was at this point I gave up and moved onto the next task'. P23
Forgetting[a]	Admit not being able to remember something; express uncertainty about recalled details	'I clicked on Education, I'm not sure if that was before or after I went to the events'. P19
Hypothesising[a]	Comments based on hypotheses rather than experience. Suggest impact problems may have on other users.	'I think teachers using this site would use it much more adeptly than I have actually used it'. P20
Task confusion	Indicate confusion or misunderstanding about interface tasks	'And I had to go back and check the question to see what it was I was actually looking for'. P23
Video cue	Responding to something they see in the video but had not noticed during task execution	'And I didn't even notice that it said galleries'. P17
Technical problems	Issues with Internet connectivity	'Then the page froze for a while'. P22

[a]Utterance categories in common with McDonald, Zhao, and Edwards (2013).

Table 4. Emergent themes from interview coding.

Category	Properties	Grouped codes	Transcript examples
Video cue	Reliance and use of the video Noticing previously unseen task-relevant information. Accounting for how or why that information was not seen.	Reliance	'Every step I took on screen refreshed my memory of what I was thinking'
		Perceived focus of attention	'I was really only looking where the mouse had been'
		Noticing unseen elements	'I noticed it (unseen link) pretty much straight away when I replayed the video'
		Responsibility	'I should have spotted it much easier than I did'
Task completion	The impact of the think-aloud on how participants approached the tasks. Their pace, flow, time awareness How they felt during task performance, if they believed it changed difficulty levels	Encoding	'I suddenly tried to remember things'
		Awareness	'It made me think more clearly I was more structured about my approach'
		Approach	'It's better to concentrate on all tasks first – then talk'
		Time	'I was aware some tasks took longer than others'
Recall	What was remembered about the session. The perceived accuracy of the data produced, task position.	Ease of Recall	'you do forget what you have actually done'
		Task-specific Recall	'I knew which task I was reporting on with the periodic one'
Verbal performance	The perceived nature of the utterances produced.	Procedures	'A lot of the time I felt I was just saying I did this, I went there'
		Difficulties	'I was highlighting difficulties. I was looking for something that was relevant to say'

Downloaded by [University of Sunderland] at 08:33 12 May 2016

become familiar with the content. In the coding process the authors read through the transcript together line-by-line, highlighting individual concepts and applying tentative labels. In order to keep the content in the words of the users, the relevant phrases were extracted and copied onto index cards with the tentative labels identified in pencil. The first author then followed this process independently to code the remaining transcripts. The result of this process was a set of index cards with initial labels applied to the constructs. The authors then met over a period of several days to further group and sort data into emergent themes. Working together the authors performed a card sort in order to confirm the codes and to categorise the data. One-by-one, each card was taken and read in turn; where two notes appeared to be about the same theme they were placed together. If a card introduced a new theme it was set apart from the rest. Category names were given to each group. Table 4 presents the high-level category names, the subgroup codes and illustrative comments from transcripts.

3. Results

We now present our results in the following order: task performance measures, utterance data and interview data. Where statistical significance is reported we use the .05 threshold.

3.1. Task performance measures

Table 5 presents the mean and standard deviation values for all task performance data.

Related samples *t*-tests revealed significant differences between the two think-aloud placements for the following measures: time on task ($t(23) = 2.76$, $p = .01$, $r = 0.5$), participants completed tasks significantly faster in the Post-session RTA condition than in the Post-task RTA condition; mouse clicks: ($t(23) = 2.48$, $p = .02$, $r = .46$), participants made significantly fewer mouse clicks in the Post-session RTA condition than in the Post-task RTA condition; divergences: ($t(23) = 2.15$, $p = .04$, $r = .41$), participants made significantly fewer divergences from the route to each answer in the Post-session RTA condition than in the Post-task RTA condition. There were no other significant differences.

3.2. Utterance data

We present the results of our utterance analysis. Non-parametric analyses were used as the data were not normally distributed.

3.2.1. The number of utterances

Figure 1 shows the percentage of each utterance category in the data as a whole and for each think-aloud placement. The largest category of utterance produced was procedural descriptions, accounting for 35% of the data followed by action explanations (20%), negative user experience (16%) and expectations (11%). One of the smallest categories of utterance was recommendations, accounting for less than 1% of the total utterances made.

Table 5. Task performance data for the two think-aloud placements.

	Post-task RTA		Post-session RTA	
	Mean	Standard deviation	Mean	Standard deviation
Number of Successful tasks	3.88	1.33	3.58	1.44
Time (in seconds) on task*	144.07	31.83	121.00	31.86
Mouse clicks*	12.31	5.30	8.88	4.54
Number of slips	3.17	3.02	2.55	2.03
Number of divergences*	5.08	6.12	1.95	3.08

*Significant difference obtained $p < .05$.

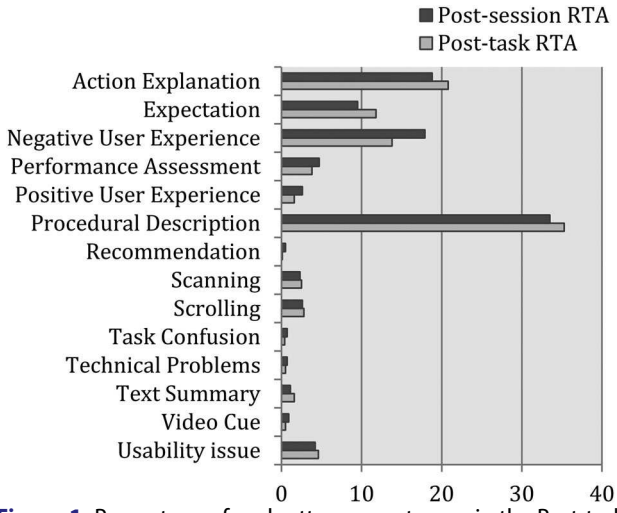

Figure 1. Percentage of each utterance category in the Post-task RTA and Post-Session RTA.

Proportionally, similar quantities of utterances were identified for each think-aloud placement. Table 6 presents the median and the lower and upper quartiles for the number of utterances made in each category for each think-aloud. A Wilcoxon Signed Ranks test revealed no difference in the total number of utterances made between the two think-aloud conditions. However, for the following three categories (those marked with an asterisk in Table 6), the Post-task RTA yielded significantly more utterances than the Post-session RTA: action explanation ($Z = -2.01$, $p = .05$, $r = -0.41$); expectation ($Z = -2.48$, $p = .01$, $r = -0.51$); procedural descriptions ($Z = -2.20$, $p = .03$, $r = -0.45$).

3.3. Participants' subjective assessment of their think-aloud

Table 7 presents the median and lower and upper quartiles for the three Likert statements about participants' think-aloud performance. The scale ranged from 1 Strongly Disagree to 5 Strongly Agree. A Wilcoxon signed ranks test revealed a significant difference between the two think-aloud placement in terms of how accurate participants believed their think-aloud to have been ($Z = 2.25$, $p < .01$, $r = 0.68$). Participants believed their verbalisations were more accurate in the Post-task RTA than in the Post-session RTA. There were no other significant differences.

3.4. Interview data: think-aloud experience

After the evaluation was complete participants helped us to understand their experiences with the think-aloud approaches through a brief semi-structured interview. Specifically, we asked participants about:

- Which of the two approaches they preferred and the reasons behind the preferences
- The impact of the approaches on task performance
- The impact and utility of the video cue

We will now discuss the themes that emerged from the interview data.

3.4.1. Think-aloud preferences

Overall, seven participants preferred the Post-session RTA; 15 preferred the Post-task RTA; 1 participant expressed no preference and one indicated that they would prefer to do the tasks and talk at the same time. From those who indicated a preference for the Post-session RTA, five participants highlighted that it was simpler and faster for them to complete all of the tasks together. All of the participants who indicated a preference for the Post-task RTA highlighted the ease of recall as the primary reason behind their preference.

Table 6. Utterances in each category for the two think-aloud placements.

	Post-task RTA			Post-session RTA		
Utterance categories	Median	Lower quartile	Upper quartile	Median	Lower quartile	Upper quartile
Action explanation*	12.50	6.25	16.00	8.50	6.25	12.75
Expectation*	6.00	3.00	9.00	4.00	2.00	8.00
Forgetting	1.00	0.00	3.00	1.50	0.00	3.00
Hypothesising	0.00	0.00	0.00	0.00	0.00	0.00
Negative user experience	7.50	5.00	10.75	7.00	3.25	14.00
Performance assessment	1.50	1.00	2.75	2.00	1.00	3.75
Positive user experience	0.50	0.00	2.75	1.00	0.00	2.00
Procedural description*	19.50	16.00	23.00	16.5	13.00	22.50
Recommendation	0.00	0.00	0.00	0.00	0.00	0.00
Scanning	1.00	0.00	2.75	1.00	0.00	2.00
Scrolling	0.00	0.00	1.75	0.00	0.00	0.75
Task confusion	0.00	0.00	0.00	0.00	0.00	0.75
Technical Problems	0.00	0.00	0.75	0.00	0.00	0.00
Text summary	1.00	0.00	1.00	0.00	0.00	1.00
Video cue	0.00	0.00	0.75	0.00	0.00	1.00
Usability issue	2.00	1.00	4.75	1.00	0.00	4.75

*Significant difference obtained $p < .05$.

Downloaded by [University of Sunderland] at 08:33 12 May 2016

Table 7. Participants self-reported assessment of their think-aloud performance.

	Post-session RTA			Post-Task RTA		
	Median	Lower quartile	Upper quartile	Median	Lower quartile	Upper quartile
The content of my think-aloud was accurate*	3.00	3.00	4.00	4.00	3.00	4.00
I relied heavily on the video replay while thinking-aloud	4.00	3.35	5.00	4.00	3.00	4.00
I remembered all of the tasks	4.00	3.00	4.75	4.00	3.25	4.00

*Significant difference obtained $p < .05$.

3.4.2. Impact of video cue

All of the participants indicated that the use of the video replay was beneficial in terms of helping them recall task-based activity and served to allay any fears they had about their ability to provide the think-aloud. However, in considering the use of the video within the interviews, a number of themes emerged.

Reliance: As indicated above, all of the participants, and regardless of think-aloud placement, highlighted their reliance on the video cue to help them recall task performance. Indeed, all 24 participants reported that the presence of the cue made the prospect of the think-aloud less daunting. For example, 'I wasn't daunted because you have the replay; without that it would have been a different prospect' P20. One participant commented that the video replay also helped her to remember tasks she felt she might otherwise have forgotten. 'If I hadn't had the replay I would have forgotten the first couple of tasks as I thought they were quite simple' P1.

Perceived focus: 23 out of 24 participants indicated that when watching the replay they focused on what they were doing rather than the broader context of the site. For example, 'I was only looking where the mouse had been' P3. One participant suggested that she took a broader view, noticing things of interest that were unrelated to the commentary she was providing. 'I wasn't looking for alternative options for the task but I did notice interesting things (events) that were going on when watching the video' P23. The participant later commented that things caught her eye during a section of the replay where the watched activity was reading and so there was little to comment on at the time. However, despite participants commenting that while watching the replay their focus was on their own activities, three comments were made in relation to the Post-task RTA that suggest that participants were taking a wider view at times. Three participants commented that they noticed items during the replay of a task that was subsequently helpful in another task. 'I noticed there was a museum objects one which I didn't notice during that task and so I used it later on' P9. Another commented: 'There were things that I was conscious of them existing when I was moving onto another task so I could probably find things easier because I had noticed them in the video of the last task' P5.

Noticing unseen elements: 14 participants commented that, regardless of the think-aloud placement, one impact of watching the replay was that they frequently noticed things they had missed during task performance. This was reported as happening for both test products. For example, 'I saw straight away the climate change wall and thought OK they were right in front of me' P19. In a similar vein another participant commented: 'the science night bit was there plain as day and I completely missed it, whereas when I was actually looking I didn't see it at all' P24. 'I saw the answer to task three during the video of task four' P10.

Responsibility: when asked about why they believed they missed certain elements during task execution, some took a pragmatic, accepting view; for example, 'It wasn't in my train of thought at the time' P19. Conversely, four participants reflected in a more negative way about their task performance, assuming responsibility for errors. For example, 'the structure of the site seemed to stay pretty much the same so there's no reason why I shouldn't have found that quicker' P8. 'I should have just looked a bit more' P23. However, 9 out of the 14 participants, who had identified that the video laid bare missed options, suggested that the reason for having missed these items was more to do with the properties of the test products. For example, layout, 'the information was hard to see as the layout was jumbled' P2; the need to read text in one task, 'I just couldn't read through that much information' P10, 'no one looks over to the right' P16; navigation, *the site wasn't intuitive*, P6.

3.4.3. Task completion

A number of themes emerged in relation to how the think-aloud placement affects task performance.

Encoding: Three participants reported becoming conscious of the need to remember their activities and thoughts during the Post-task RTA condition; 'While I was searching I had in the back of my mind that I needed to remember what I was doing' P22. Another participant commented that in the Post-task RTA 'I tried to remember things where as I didn't before (referring to Post-session RTA). I would do something then think oh I need to remember that was the reason why I did that' P14. Participants made no comments about thinking about the need to verbalise after the Post-session RTA, despite knowing this would happen before they started the tasks.

Downloaded by [University of Sunderland] at 08:33 12 May 2016

Downloaded by [University of Sunderland] at 08:33 12 May 2016

Approach: Six participants reported that the immediacy of recall in the Post-task RTA condition made them take a more relaxed approach to task performance: 'I felt more relaxed doing it that way because it was kind of block format where you would do one task and talk about it and it was done' P23. However, four participants reported that it was disruptive and preferred to just focus on the tasks and then think-aloud at the end of the session. For example, 'I found it easier to continue my train of thought and keep going from one task to another rather than break things up' P19.

Six participants commented that they engaged in more exploratory behaviours during the Post-task RTA: 'I spent a bit more time actually looking around, kind of going to where I thought it would actually be' P11. Another participant commented: 'I was trying to think of what I could do to give me something to say so that I wasn't sitting in silence' P2. One participant believed that their explorations during the Post-task RTA may have increased their session length: 'I was doing more and trying to think of more things to talk about. I think that's why I might have took longer' P12. Ten participants indicated that they just focused on the task and did not think about the need to think-aloud (for either the Post-task or Post-session RTA). For example, one participant commented: 'I just focused on the task and didn't think about what I was going to say to you' P8

Awareness: Four participants commented that the immediacy of the think-aloud in the post-task RTA made them more aware of the task-solving behaviours. 'It made me think more clearly' P12; 'I was more self-aware' P7; 'I was more aware of thinking about what I was doing' P20.

Time awareness: The interview data suggested that time became important to participants in respect of the overall session length. When reflecting on the Post-task RTA one participant commented. 'I tried to work faster so I would have less to say' P4. In relation to the Post-session RTA two participants commented that they felt the need to work faster, presumably to limit the session length: 'I felt a bit time constrained' P2.

3.4.4. Test session recall

Three themes emerged in relation to participants' memory of the test sessions: ease of recall, quality of the recalled information and test session length

Ease of recall: 20 participants commented that during recall they found reporting in the Post-task RTA condition to be easier: 'It is more fresh in your mind … you can talk about it easier … when you have one task to focus on' P1. Participants 7 and 21 commented: 'I think you can recall more information a lot easier on the local one, it was a lot easier to remember what you were doing' P7; 'It is a lot simpler because you haven't got that gap and there isn't that many things being able to interrupt your memory and lose your train of thought' P21.

Task-specific recall: In terms of recalling task performance, 15 participants indicated that they were conscious of making some generalisations in the Post-session RTA condition. For example, P12 commented 'I couldn't remember when I started one question and the other finished' and P19 commented 'it became difficult to recall exactly which task I was describing'. Ten participants suggested that the relative position of tasks within the set seemed to be of importance, in the Post-session RTA. One participant commented 'The later stages were more taxing to recall' P4. Conversely, 15 participants indicated the format of the post-task session helped participants avoid this issue.

Ten participants commented that their memory for task activity was poor during the Post-session RTA condition: 'You do forget what you have actually done' P20 and 'I couldn't remember what I was thinking' P6; 'I wasn't really sure if I would remember everything at the end' P3.

3.4.5. Think-aloud content

When asked about the content and nature of their think-aloud protocol two basic themes emerged and these were equally reported in both think-aloud conditions. Fourteen participants indicated that their think-aloud reflected both process and difficulties. Six participants indicated that they felt they talk more about the steps involved. The remaining four participants indicated that they tried to focus on difficulties because they assumed that is what we would want to know about.

Procedures: Participants highlighted that their protocol primarily reflected the steps of what they were doing. For example, 'a lot of the time I felt like I was just saying and I did that, then I went there' P7; 'I was recalling the steps more than anything else' P13.

Difficulties: Participants highlighted that during their protocol they were trying to relay the difficulties they had encountered during task performance. For example, participant 1 commented 'I was making you aware of the difficulties I had encountered'. Another commented: 'I was talking about problems I found' P20. Participant 16 commented, 'I remembered the annoyances more than anything else'.

4. Discussion

Taken together, our results suggest that the placement of the RTA affects task performance. Tasks were completed

faster, with fewer mouse clicks and fewer divergences in the Post-session RTA than in the Post-task RTA; therefore we reject H1: think-aloud placement will have no effect on task performance measures.

Turning to our analysis of participant utterances we found that the Post-task RTA yielded a greater number of procedural descriptions than the Post-session RTA, suggesting that participants had a better memory about the specifics of individual tasks. Therefore, we accept H2: the Post-task RTA will lead to an increase in the number of utterances made about task behaviours over the Post-session RTA. Moreover, participants also produced more utterances in the categories of action explanation, and expectation in the Post-task RTA condition. However, no differences were found in the number of utterances relating to the user experience. Therefore, we may partially accept H3: the Post-task RTA will lead to an increase in the number of utterances made about the user experience, expectations and explanations of behaviour over the Post-session RTA.

We now discuss our task performance and utterance data findings in detail. In so doing, we will refer to the interview data to help us consider these differences and their implications for usability testing practice.

4.1. Task performance

Although there was no difference between the two conditions for the number of successfully completed tasks, participants completed tasks significantly faster in the Post-session RTA condition; they also made significantly fewer mouse clicks and divergences from the task solutions. Previous studies that have compared two or more variants of the RTA (e.g. video-cued and eye-cued) have found no difference in task performance measures (Eger et al. 2007; Elling, Lentz, and de Jong 2011). So what might account for the difference in performance measures between the Post-task RTA and Post-session RTA?

As part of the task instructions we told participants, regardless of condition, that they would be asked to think-aloud with the help of a video cue. This knowledge of the impending need to think-aloud may have influenced behaviour in a number of possible ways.

This knowledge may have served to increase participants' cognitive load in that they needed to focus on task completion and they may have felt the need to actively remember what they were doing. Block, Hancock, and Zakay (2010) identify a number of different ways in which cognitive load might be manipulated; one way that is pertinent to the present study is to instruct participants to try and remember information for a later test while they are carrying out some other test-related activity. Therefore, situations of high cognitive load arise when participants are instructed about the need to remember information and cognitive load is low when no such instructions are present. While we did not ask users to try and remember their activity, we did tell them in both conditions that they would be asked to provide a think-aloud at a later point, accompanied by a video replay. Therefore, our participants may have had increased cognitive load over an RTA in which users are not told about the need to think-aloud until after task completion.

Arguably, one might have expected the perceived or actual cognitive load to be greater for the Post-session RTA because they had to recall a larger number of tasks; therefore one might expect that their task performance might have suffered. However, this was not the case; participants were faster, made fewer clicks and divergences during the Post-session RTA than the Post-task RTA. It may be, therefore, that the greater distance between action and recall in the Post-session RTA served to reduce the impact or participants' awareness of the need to think-aloud. However, in the Post-task RTA, the gap between action and recall was shorter, meaning that the requirement to think-aloud may have been more pressing. Consequently, the need to remember activity may have diverted cognitive resources away from task performance, which had a deleterious impact on task execution. Indeed, in the post-test interview, some participants commented that during the Post-task RTA condition they were making a conscious effort to remember what they were doing as they knew they would be asked to speak about it straight away 'I was more aware of thinking about what I was doing because I was thinking I'm going to have to remember to tell you about it' P17. However, these comments were confined to only three participants. Moreover, participants indicated that they were not daunted at the prospect of thinking-aloud, regardless of condition. Alternative explanations may, therefore, be possible.

The difference in task performance may be due to a disruption of task flow. The Post-task RTA caused a break between tasks; it may be that the necessity to re-orient after each task effectively slowed down task completion. Indeed, four users did highlight this as an issue and indicated a preference for Post-session RTA on that basis.

Finally, the Post-task RTA may have influenced participants' expectations about the content of their think-aloud. Indeed, in this paper we suggest that the proximity between action and recall in the Post-task RTA condition means that we, as evaluators, might reasonably expect users' comments to be more detailed or insightful.

The users themselves, during the Post-task RTA, may have also formulated such expectations about their own

Downloaded by [University of Sunderland] at 08:33 12 May 2016

Downloaded by [University of Sunderland] at 08:33 12 May 2016

performance. Therefore, it may be that participants felt the pressure to have more to say. Researchers have alluded to the social impact of think-aloud. For example, McDonald and Petrie (2013) who compared different variants of the concurrent think-aloud to silent working report that their participants made comments to indicate that impression management during the concurrent think-aloud was a factor for them. Indeed, one of our own participants commented: 'I was trying to think of what I could do to give me something to say so that I wasn't sitting in silence'.

Clearly, the difference in performance has generated more questions than answers and further work is required to fully understand the reasons for these differences in tasks' performance between the two conditions. Although not available to us at the time, eye movement data have been found to be particularly helpful in this respect (see e.g. Hertzum, Hansen, and Andersen 2009; Gerjets, Kammerer, and Werner 2011).

4.2. Utterance data

Turning to our analysis of the utterance data, overall, we found that there were no differences in the number of utterances made between the two conditions. There were also no differences in the categories of utterance that were unique to either the Post-session RTA or the Post-task RTA conditions, and each condition contained a similar proportion of each utterance type. Regardless of think-aloud placement, the most populous categories of utterance were: procedural description, action explanation, negative user experience and expectation. This finding adds support to Bowers and Snyder (1990), Page and Rahimi (1995), van den Haak, de Jong, and Schellens (2003) and Guan et al. (2006) who found that the RTA method as a whole generates explanatory data. Overall, participants generated very few recommendations; this finding is similar to studies investigating the concurrent method (see e.g. Zhao and McDonald 2010) and the use of both concurrent and retrospective reports (see e.g. McDonald, Zhao, and Edwards 2013). Taken together, these findings suggest that, by and large, participants rarely express recommendations for design changes during think-aloud studies.

The Post-task RTA yielded significantly more utterances of the following types: action explanation, expectation and procedural description. It would seem, therefore, that during the Post-task RTA participants provided more detailed descriptions of what they were doing and also their reasons behind their actions. Moreover, they also described when and how their expectations were, or were not, met by the test products. The greater number of these types of utterances lends support to the argument that immediate task review would provide more detailed insights because the information is fresh in the user's mind (Frøkjær and Hornbæk 2005; Følstad and Hornbæk 2010).

We also found instances of forgetting and hypothesising but overall they were low in number (3% of the total data set). These findings support Guan et al. (2006); McDonald, Zhao, and Edwards (2013) who also found a small number of utterances that would suggest inaccurate recall. These results suggest that RTA, in general, is valid and that the test session replay serves as a durable memory cue, particularly for the Post-session RTA (Ohnemus and Biers 1993).

Comparing our utterance analysis with a recent investigation of the utterance content of a Post-session RTA by McDonald, Zhao, and Edwards (2013) we found a greater number of procedural descriptions; approximately 35% of the total utterances made reflected task-based behaviours as opposed to around 11% in their study. Video replay undoubtedly cues the production of task-based utterances and this difference is therefore to be expected. We also note an increase in the number of explanations about the reasons for actions (approximately 20% of utterances) and users' expectations (approximately 11% of utterances) compared to 6% and 3% in McDonald et al.'s study, suggesting that, overall, the video cue may have had a positive impact in the elicitation process. However, McDonald, Zhao, and Edwards (2013) report significantly more utterances in relation to usability issues; 35% of the utterances made in their RTA were about usability issues compared to only 4% in our study. We believe the differences may be due to the following factors: first, in preparing the task set for their study McDonald, Zhao and Edwards first conducted an expert inspection of the product and used this to inform the task set; therefore we might reasonably expect that more verbalisations about usability issues might result. In the present study, we needed to use two products to control for practice effects. Therefore, our focus in task derivation was to ensure that tasks were matched and of equal difficulty; we did not base the tasks around inspection conjectures. Second, their RTA followed a concurrent think-aloud; it is therefore possible that the concurrent verbalizations primed the RTA, reinforcing issues in the users' minds meaning that they were more likely to be reported during the Post-session RTA.

4.3. Implications for usability practice

In common with the concurrent technique, there are a number of modifications that can be made to the way in which we elicit the RTA. In this section, we consider

the implications of the findings of this study for usability practice. Specifically, we consider the implications for the use of a video cue, task considerations and session length.

4.3.1. Video-cued RTA

The use of a video replay to cue the RTA is a double-edged sword. On the one hand, the video may help increase the accuracy of the verbal data (Guan et al. 2006) and the participants in our study indicated that they relied upon it greatly. However, the use of video replay does serve to double the length of the think-aloud session, which may be problematic for practitioners who may be constrained by both time and resources.

One general concern in using a video cue in either Post-task RTA or Post-session RTA is that participants might respond to the cue rather than re-telling their task-based activities (Leow 2002; Cotton and Gresty 2006). In the present study, we did find instances of utterances about the video cue but this category contained instances of participants spotting links or menu options they had not seen during task execution, and as such the category was helpful in identifying the reasons behind some of divergences made during task execution.

McDonald, Zhao, and Edwards (2013) report participants providing additional comments about features of their test product that did not relate directly to task performance during their RTA. No such observations were made in the present study. However, the pace of the RTA in McDonald, Zhao and Edwards's, study was not controlled by an external agent (the video replay) as was the case in the present study. Moreover, because their participants were still able to interact with the product they were free to engage in a more active manner.

A further issue in respect of using the Post-task RTA is that while watching a replay of a given task, users might notice interface elements that would be of help in subsequent tasks. Indeed, we found some evidence of this activity within our own study. During the interview, some of our participants revealed that they had noticed links that would prove useful in subsequent tasks. However, we do not know (without support from eye tracking data) for how many other participants this was also true. Therefore, there is the potential for task-based video-cued review to, inadvertently, affect subsequent task performance and potential users of this approach need to be cognisant of this potential source of bias.

4.3.2. Task considerations and session length

Our study used only six tasks per condition and in the Post-session RTA, some participants indicated difficulties in recalling task-based activities, even with the video replay. In particular, they encountered difficulties in identifying one task from another. We endeavoured to support differentiation by using a varied task set; however, in the context of usability testing where evaluators are under pressure to learn as much as possible in a short space of time (Chilana, Wobbrock, and Ko 2010), task sets may be extended. Therefore, the issue of task differentiation will be a perennial problem for the Post-session RTA. The Post-task RTA overcomes this issue in that participants are able to focus on one task at a time, and in comparing the two alternative placements of the think-aloud it is this aspect of the Post-task RTA that participants preferred and it appears to be why they believed their Post-task RTA to be more accurate than the Post-session RTA. It may be therefore that practitioners could consider Post-task RTA in situations in which task differentiation is hard to achieve. However, differences in the task-based performance data require further consideration. It may be, therefore, that Post-task RTA should be used with caution and perhaps only once within a session, with the instruction only being given after the task is complete.

4.4. Limitations and future work

The study presented in this paper has a number of limitations. First, the first author functioned as both the test facilitator and the primary data coder. In an ideal situation separate individuals would have performed these activities. To mitigate the potential bias this introduced the following measures were taken: (i) there was a delay of three weeks between data collection and transcription and a further five weeks for subsequent qualitative analysis; (ii) the second author independently coded a sub-set of the data and cross-checked all of the remaining qualitative data without knowledge of which TA placement condition the data came from.

Second, while we report a range of task performance measures and have attempted to explain our findings in light of participant feedback during interviews, we cannot pinpoint further behavioural differences between the two conditions during task performance. Other researchers have used eye tracking to explore differences in users' attention resources during task performance with different variants of the concurrent think-aloud (e.g. Hertzum, Hansen, and Andersen 2009; Gerjets, Kammerer, and Werner 2011). At the time of testing, we were unable to record eye tracking data which may have helped us to clarify the differences between the two conditions; however, such an investigation is now underway in our laboratory.

Downloaded by [University of Sunderland] at 08:33 12 May 2016

4. Conclusions

In their seminal work on protocol analysis Ericsson and Simon (1984, 1993) suggest the collection of both concurrent and retrospective reports. In usability testing, however, we tend to collect either concurrent or retrospective protocols. In this paper, we do not argue for the replacement of one technique with the other. Indeed we would suggest, as others have done before (e.g. Gray and Salzman 1998), that there is unlikely to be one single best approach as contextual factors such as the product, tasks types, user groups will always influence the success (or otherwise) of a given technique. Instead, we suggest that we need to investigate the elicitation procedures of individual approaches in order to discover how and when techniques might prove fruitful and what the practical ramifications of their use might be.

Researchers considering the use of the RTA are faced with a number of elicitation options, with the placement of the think-aloud being an important consideration. Placing the think-aloud after each task does appear to increase the number of explanatory utterances produced over a Post-session RTA. However, the differences in task performance give rise to concerns. Despite there being no difference in the number of successfully completed tasks, the Post-task RTA may have influenced users' task-solving strategies or increased their cognitive load and as a consequence may threaten the validity of the test. We need therefore to further understand the cause of these differences; such work is currently underway in our laboratory.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their detailed and constructive feedback on the paper; it was very much appreciated. We also thank Dr Ken McGarry for the many useful discussions we have had. Finally, thanks are owed to our test participants who gave their time freely to help us with our research.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Block, R. A., P. A. Hancock, and D. Zakay. 2010. "How Cognitive Load Affects Duration Judgments: A Meta-analytic Review." *Acta Psychologica* 134 (3): 330–343.

Boren, M. T., and J. Ramey. 2000. "Thinking Aloud: Reconciling Theory and Practice." *IEEE Transactions on Professional Communication* 43 (3): 261–278.

Bowers, V. A., and H. L. Snyder. 1990. "Concurrent versus Retrospective Verbal Protocols for Comparing Windows Usability." In *Proceedings of the Human Factors Society 34th Annual Meeting*, 1270–1274 Santa Monica, CA: HFES.

Chilana, P. K., J. O. Wobbrock, and A. J. Ko. 2010. "Understanding Usability Practices in Complex Domains." In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI'10, Atlanta, GA, USA, April 10–15, 2337–2346. New York: ACM.

Cotton, D., and K. Gresty. 2006. "Reflecting on the Think-aloud Method for Evaluating E-learning." *British Journal of Educational Technology* 37 (1): 45–54.

Eger, N., L. J. Ball, R. Stevens, and J. Dodd. 2007. "Cueing Retrospective Verbal Reports in Usability Testing through Eye-movement Replay." In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers XXI: HCI … but Not as We Know It*, HCI'07, 129–137. Swindon: British Computer Society.

Elling, S., L. Lentz, and M. de Jong. 2011. "Retrospective Think-aloud Method: Using Eye Movements as an Extra Cue for Participants' Verbalisations." In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, CHI'11, Vancouver, BC, Canada, May 7–12, 1161–1170. New York: ACM.

Ericsson, K. A., and H. A. Simon. 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge: MIT Press.

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

Fox, M. C., K. A. Ericsson, and R. Best. 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.

Følstad, A., and K. Hornbæk. 2010. "Work-domain Knowledge in Usability Evaluation: Experiences with Cooperative Usability Testing." *Journal of Systems and Software* 83 (11): 2019–2030.

Frøkjær, E., and K. Hornbæk. 2005. "Cooperative Usability Testing: Complementing Usability Tests with User-supported Interpretation Sessions." In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA'05, Portland, Oregon, USA, April 2–7, 1383–1386. New York: ACM.

Gerjets, P., Y. Kammerer, and B. Werner. 2011. "Measuring Spontaneous and Instructed Evaluation Processes During Web Search: Integrating Concurrent Think-Aloud Protocols and eye-Tracking Data." *Learning and Instruction* 21 (2): 220–231.

Gray, W. D., and M. C. Salzman. 1998. "Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods." *Human-Computer Interaction* 13 (3): 203–261.

Guan, Z., S. Lee, E. Cuddihy, and J. Ramey. 2006. "The Validity of the Stimulated Retrospective Think-aloud Method as Measured by Eye Tracking." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1253–1262. New York, NY: ACM.

van den Haak, M. J., M. de Jong, and P. J. Schellens. 2003. "Retrospective vs. Concurrent Think-aloud Protocols: Testing the Usability of an Online Library Catalogue." *Behaviour & Information Technology* 22 (5): 339–351.

van den Haak, M. J., M. D. T. de Jong, and P. J. Schellens. 2004. "Employing Think-aloud Protocols and Constructive Interaction to Test the Usability of Online Library

Downloaded by [University of Sunderland] at 08:33 12 May 2016

Catalogues: A Methodological Comparison." *Interacting with Computers* 16 (6): 1153–1170.

van den Haak, M. J., M. D. de Jong, and P. J. Schellens. 2007. "Evaluation of an Informational Web Site: Three Variants of the Think-aloud Method Compared." *Technical Communication* 54 (1): 58–71.

van den Haak, M. J., M. D. T. de Jong, and P. J. Schellens. 2009. "Evaluating Municipal Websites: A Methodological Comparison of Three Think-aloud Variants." *Government Information Quarterly* 26 (1): 193–202.

Hertzum, M., P. Borlund, and K. B. Kristoffersen. 2015. What Do Thinking-aloud Participants Say? A Comparison of Moderated and Unmoderated Usability Sessions. *International Journal of Human-Computer Interaction* 31 (9): 557–570. doi:10.1080/10447318.2015.1065691.

Hertzum, M., K. D. Hansen, and H. H. K. Andersen. 2009. "Scrutinising Usability Evaluation: Does Thinking Aloud Affect Behaviour and Mental Workload?" *Behaviour & Information Technology* 28 (2): 165–181.

Hertzum, M., and K. D. Holmegaard. 2013. "Thinking Aloud in the Presence of Interruptions and Time Constraints." *International Journal of Human-Computer Interaction* 29: 351–364.

Leow, R. P. 2002. "Models, Attention and Awareness in Sla." *Studies in Second Language Acquisition* 24 (1): 113–119.

McDonald, S., H. Edwards, and T. Zhao. 2012. "Exploring Think-alouds in Usability Testing: An International Survey." *IEEE Transactions on Professional Communication* 55 (1): 2–19.

McDonald, S., and H. Petrie. 2013. "The Effect of Global Instructions on Think-aloud Testing." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, April 27–May 2, 2941–2944. New York, NY: ACM.

McDonald, S., T. Zhao, and H. M. Edwards. 2013. Dual Verbal Elicitation: The Complementary use of Concurrent and Retrospective Reporting within a Usability Test. *International Journal of Human Computer Interaction*, 29: 647–660.

McDonald, S., T. Zhao, and H. M. Edwards. 2015. "Look Who's Talking: Evaluating the Utility of Interventions During an Interactive Think-aloud." *Interacting with Computers* 28: 387–403. doi:10.1093/iwc/iwv014.

Nørgaard, M., and K. Hornbæk. 2006. "What Do Usability Evaluators Do in Practice? An Explorative Study of Think-aloud Testing." In *Proceedings of the 6th Conference on Designing Interactive Systems*, University Park, Pennsylvania, June 26–28, 209–218. New York: ACM Press.

Ohnemus, K. R., and D. W. Biers. 1993. "Retrospective Versus Concurrent Thinking-out-Loud in Usability Testing." *Proceedings of Human Factors and Ergonomics Society Annual Meeting* 37 (17): 1127–1131.

Olmsted-Hawala, E. L., E. D. Murphy, S. Hawala, and K. T. 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 28th International Conference on Human Factors in Computing Systems*, 2381–2390 Atlanta: ACM Press.

Page, C., and M. Rahimi. 1995. "Concurrent and Retrospective Verbal Protocols in Usability Testing: Is there Value Added in Collecting Both?" *Proceedings of Human Factors and Ergonomics Society Annual Meeting* 39 (4): 223–227.

Shi, Q. 2008. "A Field Study of the Relationship and Communication between Chinese Evaluators and Users in Thinking Aloud Usability Tests." In *Proceedings of the 5th Nordic Conference on Human–computer Interaction: Building Bridges*, Lund, Sweden, October 20–22, 344–352. New York: ACM Press.

Taylor, K. L., and J. P. Dionne. 2000. "Accessing Problem-solving Strategy Knowledge: The Complementary use of Concurrent Verbal Protocols and Retrospective Debriefing." *Journal of Educational Psychology* 92 (3): 413–425.

Yang, S. C. 2003. "Reconceptualising Think-aloud Methodology: Refining the Encoding and Categorising Techniques via Contextualised Perspectives." *Computers in Human Behaviour* 19 (1): 95–115.

Zhao, T., and S. 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*, NordicCHI'10, Reykjavik, Iceland, October 16–20, 581–590. New York: ACM.

Zhao, T., S. McDonald, and H. M. Edwards. 2014. "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. doi:10.1080/0144929X.2012.708786.

Downloaded by [University of Sunderland] at 08:33 12 May 2016