This is a repository copy of *Comparing Concurrent and Retrospective Verbal Protocols for Blind and Sighted Users*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/109260/

Version: Published Version

Proceedings Paper:
Savva, Andreas, Petrie, Helen orcid.org/0000-0002-0100-9846 and Power, Christopher Douglas orcid.org/0000-0001-9486-8043 (2015) Comparing Concurrent and Retrospective Verbal Protocols for Blind and Sighted Users. In: Human-Computer Interaction. , pp. 55-71.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Comparing Concurrent and Retrospective Verbal Protocols for Blind and Sighted Users

Andreas Savva(✉), Helen Petrie, and Christopher Power

Human Computer Interaction Research Group, Department of Computer Science, University of York, York YO10 5GH, UK
{asl5l7, helen.petrie, christopher.power}@york.ac.uk

Abstract. Verbal protocols are widely used in user studies for evaluating websites. This study investigated the effectiveness and efficiency of concurrent and retrospective verbal protocols (CVP and RVP) for both blind and sighted participants, as well as participant workload and attitudes towards these methods. Eight blind and eight sighted participants undertook both protocols in a website evaluation. RVP was more effective as measured by problems encountered for both groups, although it was no more efficient than CVP. The severity of problems identified by both protocols was equivalent. As measured on the NASA TLX, participants found RVP found more demanding than CVP. Sighted participants found rating problems during CVP more disruptive than blind participants. These results show that RVP is a more useful protocol for practitioners and researchers even though it takes more time and is more demanding for participants. It is equally applicable for both blind and sighted participants.

Keywords: User evaluation · Think aloud protocol · Concurrent verbal protocol · Retrospective verbal protocol · Web accessibility · Web usability · Blind users

1 Introduction

In user-based studies to evaluate websites, participants typically "think aloud" while undertaking tasks to identify problems. The thinking aloud may be performed concurrently with conducting the task, known as a concurrent verbal protocol (CVP), or retrospectively while reviewing recordings of their performance on a task, known as a retrospective verbal protocol (RVP). A number of studies have compared these two types of verbal protocol with sighted participants, in terms of the information gathered [3, 17] and the number of problems revealed [25, 26]. However, a comparison of these protocols when used with blind participants has not yet been performed, in spite of the fact that there are a number of studies which have used verbal protocols with blind participants [8, 12, 19, 20, 22]. CVP may add additional effort particularly for blind participants, as the mental effort of using the web for blind users with screen readers is typically greater than understanding the web visually. This is because blind users need to recall all the keyboard commands that they use to interact with the web, whereas sighted users can rely on recognition of icons and menu items if need be.

© IFIP International Federation for Information Processing 2015
J. Abascal et al. (Eds.): INTERACT 2015, Part I, LNCS 9296, pp. 55–71, 2015.
DOI: 10.1007/978-3-319-22701-6_5

However, few studies have compared the two verbal protocols in terms of the workload they place on participants. In addition, there are no studies comparing the two protocols with blind participants in terms of information gathered, problems revealed and the workload of the protocol. As blind participants are the most common disabled user group to participate in evaluations of websites for their accessibility, research is needed to establish which protocol is better to use.

We conducted a study with blind and sighted participants, performing both CVP and RVP, to compare the two protocols in terms of effectiveness, efficiency and the effect the protocols have on the two participant groups.

2 Related Work

In user-based studies of websites, a number of users who represent the target audience perform a number of tasks on the target websites. The most basic user evaluation has users performing a task in order to measure the users' performance on it. In addition, users can perform tasks while performing a verbal protocol, which can offer insight into the users' thought processes, the problems they encountered and their problem solving strategies [15]. The verbal protocol derives from the work of Ericsson and Simon, and was originally used as a research method in cognitive psychology [7]. It was introduced into the usability field by Lewis [13]. The underlying concept of this approach is the passive role of the evaluator, as there is no interaction between the evaluator and the participants while they perform the verbal protocol, except to remind them to think out loud if they become silent. Even the verbal protocol is based on this approach, some practitioners and researchers do not maintain the passive role of the evaluator [1, 16, 23].

Boren and Ramey [1] observed the verbal protocol methods used in two companies. Their results demonstrated that evaluators did not instruct participants comparably, as there were variations in instructions on how to think out loud. Moreover, most of the practitioners started immediately with the tasks, without giving participants any practice in the verbal protocol technique. Also, Boren and Ramey found inconsistencies among the prompts that evaluators used to remind participants when they fell silent for a period of time. Finally, most evaluators intervened in ways that did not reflect the approach of Ericsson and Simon. Based on these observations, Boren and Ramey [1] proposed a new approach, based on speech communication theory, in which evaluators have a more active role in comparison to the Ericsson and Simon approach.

Several studies have investigated if the change in approach affects participants' performance [9, 18]. In 2004, Krahmer and Ummelen [9] conducted a study with 10 participants, who performed a verbal protocol using either the original Ericson and Simon approach or the more active Boren and Ramey approach. They found that the approach did not affect the number of problems detected, however there was difference in participants' performance. Participants using the Boren and Ramey approach were more successful in completing tasks. Olmsted-Hawala et al. [18] compared the two approaches in a study with 80 participants. They found no differences in participants' performance between the two approaches.

As mentioned above, the verbal protocol may be performed concurrently or retrospectively. In CVP participants think out loud while doing the task, whereas in RVP participants perform the task first in silence and then think out loud while watching a video of themselves doing the task [6, 15, 21]. In the case of blind participants, they listen to the audio of themselves using the screen reader, which is the equivalent cue for them.

Several studies have been conducted to compare the differences found in the information gathered between the two methods [3, 10, 17]. Bowers and Snyder [3] conducted a study comparing the two protocols in a multiple window task. Their results revealed that during CVP more procedural information was collected, whereas during RVP more design changes and explanations were collected. Ohnemus and Biers [17] found that in RVP participants produced more statements which were useful for designers than in CVP. Kuuesela and Paul [10] compared the two protocols in terms of effectiveness for revealing human cognitive processes. Their results showed that CVP provides more insight into decision making processes, whereas RVP provides more statements about the participants' final choice.

Studies have also been conducted to compare the effects of the two protocols on participants' performance. A number of studies found that there is no difference between the two protocols in terms of task performance [17, 25, 26]. However, there are also several studies that showed that verbalization could have an impact on participants performance: either improved [30] or worsened it [29].

Uncovering user problems is one of the most important features in conducting user-based evaluations. A number of studies compared the number of problems revealed between the two protocols and most of them have demonstrated that the two protocols revealed a comparable set of problems [25–28]. However, these studies have some limitations, as only one website was used in each one of the studies. More extended research needs to be conducted to compare the two protocols in terms of the number of user problems revealed.

Some user-based evaluation studies are undertaken with disabled people to identify accessibility problems. The most frequent disabled groups involved are blind users. Studies that have included blind participants have almost exclusively had them perform CVP [8, 12, 19, 20, 22]. While it seems the standard protocol to use, it is a method that adds additional workload to the users in vocalizing their thoughts about their actions and the problems they encounter while trying to undertake a task. For blind users in particular, it is likely that the workload of the task is already high when they are working with a screen reader due to the need to remember several different modes, shortcut keys and settings. As their workload is likely to be higher than that of sighted participants, it is possible that RVP is more appropriate for blind participants and that this protocol will yield better results. No research could be found with comparing the verbal methods with any disabled user group.

Chandrashekar et al. [4] conducted a user-based evaluation study with six visually impaired participants, evaluating a website using CVP. They noted that blind participants did not respond when they were prompted using defined time intervals. Moreover, they stated that it is not feasible to have blind participants think aloud concurrently, as they use the screen reader to read the text on the page. Some participants were not willing to stop the screen reader in order to think out loud, as it

interrupted the flow of the task. Also, they noted that the participants did not offer many comments, even though the researcher prompted them. Our experience of conducting many evaluations with blind participants is that they are quite happy to mute the screen reader when they are think out loud; even if they fail to remember to do this, it is usually possible to understand both what the participant is saying and the screen reader output. However, it may well be the case that this interrupts the flow of the task more than it would for sighted participants.

Even though some variations of verbal protocols for blind participants have been proposed [2, 24], they have not being used by other researchers. Further research needs to be conducted to compare the two verbal protocols, with both blind and sighted participants.

In this paper, a user-based study with blind and sighted participants comparing the two protocols, CVP and RVP, is presented. This study addressed a number of research questions, which can be grouped into three areas:

Effectiveness of CVP versus RVP:

- Does one protocol identify more distinct problems than the other?
- Do blind and sighted participants identify the same number of problems with each protocol?
- Does one protocol identify more severe problems than the other?
- Do the two protocols identify the same problems?

Efficiency of CVP versus RVP:

- Does one protocol identify problems more rapidly?

Effect of CVP and RVP on blind and sighted participants:

- Does one protocol demand greater workload for participants, either blind or sighted?
- Does one protocol make participants more self-conscious than the other?
- Do participants prefer one method in comparison to the other?

3 Method

3.1 Design

This study was a task-based user evaluation with blind and sighted participants using two different verbal protocols, CVP and RVP. A mixed design was used with user group as the between-participant independent variable with two levels (blind or sighted participants) and the within-participant independent variable with two levels (CVP and RVP).

Participants evaluated two websites with each protocol. In addition to talking the researcher through about what they were thinking, each time a participant encountered a problem, they were asked to rate its severity on a scale from 1 (cosmetic) to 4 (catastrophic). Problems were considered everything that participant felt that was a problem, whether it was caused by the website, the browser or the screen reader. After

each session, participants were asked to complete the NASA TLX, a subjective workload questionnaire [14], as well as a questionnaire about their experience with the methods they had used.

3.2 Participants

Sixteen participants took part in the study, eight blind screen reader users and eight sighted users. Six of the blind participants were men and two were women. Ages ranged from 23 to 64 years (median = 43 years). Three of the participants were congenitally blind while the remaining five lost their sight between the ages of 26 and 49.

Sighted participants were selected to achieve as close a matched sample as possible with the blind participants on gender, age, operating system used, web experience and web expertise. Thus, six of the sighted participants were men and two were women. Ages ranged from 22 to 55 years (median = 40 years).

Participants rated their experience and expertise on the web using a five-point Likert items (1 = Very low to 5 = Very Good). The average rating for web experience for blind participants was 4, whereas for sighted participants was 4.5. On web expertise, the average rating of blind participants was 3.8, for sighted participants it was 3.6.

All blind participants used screen readers to access computers and the web for home and work. Five used JAWS (running on the Windows OS) and three used VoiceOver (running on Mac OSX). The JAWS version varied from JAWS 12.0 to JAWS 15.0 (the latter being the latest version of JAWS when the study was conducted). Participants who used VoiceOver used the latest version on the Mac OS Mavericks operating system (the latest version of Mac OS when the study was conducted). Blind participants were asked to rate their experience and expertise of using screen readers on a five-point Likert item (1 = “Very Low” to 5 = “Very Good”). The average rating for experience and expertise using screen readers was 4 and 3.9, respectively.

Six participants used Mac OSX (three blind and three sighted) and 10 participants used Windows (five blind and five sighted). The majority of the blind participants who used Windows mentioned Internet Explorer as their primary browser and all of the participants who used Mac OSX reported using Safari as their primary browser. Of the sighted participants, the ones who used Windows mentioned Chrome as their primary browser and one of them mentioned Internet Explorer. Of the ones using Mac OSX, one of them mentioned Chrome, whereas the other two mentioned Safari as their primary browser.

3.3 Equipment and Materials

For participants who use the Windows OS, the study was conducted using a desktop computer running Windows 8 with speakers, keyboard and a 2-button mouse with scroll wheel. For participants who use the Mac OSX, the study was conducted using a MacBook Pro laptop running the Mavericks Operating System, with speakers, and 2-button mouse with scroll wheel. In addition, blind participants used the version of

JAWS they were most familiar with or used the VoiceOver version that comes with Mavericks OS.

The sessions were recorded using Morae 3.1 on Windows or ScreenFlow 4.0.3 on Mac OSX. These recordings included audio, for analyzing the verbal protocols, screen activity for understanding the users' actions, and participants' facial expressions.

After each session participants completed the NASA TLX, a subjective workload questionnaire [14]. NASA TLX measures the overall effort or workload of the task, but also six different measurements of mental demand, physical demand, temporal demand, effort, frustration and performance of the participant.

At the end of the CVP session participants completed a questionnaire about the method using 5-point Likert items:

- Protocol interrupt (Q1): To what extent did thinking aloud during the task interrupt the flow of the task?
- Rating interrupt (Q2): To what extent did having to rate the problems for severity during the task interrupt the flow of the task?
- Protocol concentration (Q3): To what extent did thinking aloud during the task affect your concentration during the task?
- Rating concentration (Q4): To what extent did having to rate the problems for severity during the task affect your concentration during the task?
- Protocol real life (Q5): To what extent do you feel that thinking aloud during the task changed the way you did the tasks in comparison on how you might do it in real life?
- Protocol tiring (Q6): How tiring was it to do think aloud during the task?

Participants answered Q1 – Q5 using a scale: 1 = "Not at all" to 5 = "Very much", and Q6 using a scale: 1 = "Not at all tiring" to 5 = "Very tiring".

At the end of both verbal protocols participants were asked to complete the following question:

- To what extent did thinking aloud during the task/replay of the task made you self-conscious about what you were doing?

Participants answered this question using a scale: 1 = "Not at all" to 5 = "Very much".

Finally at the end of the session, participants were asked to select which one of the two verbal protocols they preferred conducting and to explain why they chose that preference.

3.4 Websites and Tasks

Four websites from different domains were used: a government website (www.gov.uk), a real estate website (www.rightmove.co.uk), an ecommerce website (www.boots.com) and a news website (www.channel4.com).

The tasks used were:

- Gov.uk: Find how much it is going to cost to arrange a meeting to apply for a National Insurance number from your mobile phone number.

- Rightmove: Find a house to rent with a minimum of two bedrooms and a rent of no more than £1200 per month, near to a secondary school (a postcode was provided).
- Boots: Find the cheapest, five-star rated car seat for a two-year old child who weights 24 kg.
- Channel4: Find which movie will be on Film4 at 9 pm the day after tomorrow.

The tasks that were used investigate different design aspects of the websites, such as information architecture, navigation, content, headings, links, images, forms and tables. Tasks were undertaken by the first author using JAWS 15.0 and VoiceOver, to check that it was possible for screen reader users to be able to complete the tasks.

3.5 Procedure

The study took place in the Interaction Laboratory at the Department of Computer Science of the University of York and at the National Council For the Blind of Ireland (NCBI) in Dublin. Participants were first briefed about the study and were asked to sign an informed consent form. In order to avoid any conflicts between the technology and participants' preferences, participants were asked which browser they would like to use. Blind participants were also asked which screen reader they preferred and which version. They were also given the option to adjust the computer display, sound and related software to their preferences in order to match to their usual setup.

The researcher gave a demonstration on how to perform the verbal protocol the participant was about to conduct. Participants tried the protocol out using a practice website not analysed in the study.

When participants were comfortable doing the appropriate verbal protocol, they were asked to perform each task. Depending on which protocol participants were using, they performed CVP or RVP. The verbal protocol approach that was used was based on the Boren and Ramey [1] approach. During the CVP condition they thought out loud as they performed the tasks. When participants were quiet for an extended period of time, they were prompted with "What are you thinking about?" to remind them to vocalize their thoughts. No predetermined time intervals were used to remind blind participants, as there were occasions when blind participants were silent for a long time because they were clearly listening to the text from the website using the screen reader. Thus, the use of reminding prompts relied on researcher's discretion. When participants encountered a problem, however minor, the researcher asked them to describe the problem and rate its severity using a four-point scale. The rating scale is based on Nielsen's severity ratings for usability [15]. However the description of the problem was adapted to a user-centred description, as follows:

- Cosmetic problem (1): This problem on the website is making it slightly difficult to complete my task
- Minor problem (2): This problem on the website is making it difficult to complete my task
- Major problem (3): This problem on the website is making it very difficult to complete my task
- Catastrophic problem (4): This problem on the website makes my task impossible to complete

During the RVP condition participants performed the task in silence, then they reviewed the task as the video (or for the blind participants, the audio) of the task was played back. Participants controlled the video/audio using the spacebar button of the computer to pause and resume the flow, in order to think out loud. Similar prompting and problem severity rating procedures were used in the RVP conditions as in the CVP.

This procedure was repeated for each website. After doing two websites with one protocol participants were asked to complete the NASA TLX and the questionnaire about the method they used. The procedure was then repeated for the second verbal protocol.

After completing both protocols, participants were asked to choose which one of the two protocols they preferred and to explain why, as well as to complete a demographic questionnaire. Finally, participants were debriefed about the study and the researcher answered their questions.

3.6 Data Analysis

The video recordings of each participant were reviewed, in order to code the problems and perform a problem matching technique. In the first phase of analysis, the problems identified by the users were structured using a variation of the model of Lavery et al. [11], in which the problems are analysed in relation to four components: cause, breakdown, outcome and design change. For this study, the design change component was not used. The second phase of analysis involved identifying distinct problems. Problem instances checked if there were distinct problems, that is a problem that may have been encountered by more than one participant or by the same participant on more than one occasion on the same website in the same context.

In order to check the validity of the analysis, inter-coder reliability was performed by another researcher of the Human Computer Interaction Research Group on a sample of the data. This yielded an agreement of more than 90 % on both phases of the analysis.

For this analysis we concentrate only on the number of problems and their severity rating, not the different causes or different types of the problems.

4 Results

A total of 260 instances of problems yielded 136 distinct problems were identified, across both protocols and both user groups. The average number of instances of problems was 8.13 per participant per website.

To investigate whether one protocol identified more problem instances than the other and whether blind or sighted participants identified more problem instances, a 2-way mixed ANOVA was conducted on the number of problem instances identified in each protocol condition and by blind and sighted participants. The analysis revealed a significant main effect for protocol ($F = 6.93$, $df = 1,14$, $p < 0.05$). The mean number of problem instances identified using CVP was 6.56 (SD = 2.39), whereas in RVP it was 9.69 (SD = 4.27). There was no significant main effect for user group ($F = 3.06$,

df = 1,14, n.s.). Thus, there was no difference between blind and sighted participants in the number of problem instances identified. Finally, there was no interaction between protocol and user group (F = 0.00, df = 1,14, n.s.).

To investigate the severity of problem instances identified in the two protocols and by blind and sighted participants, a 2-way mixed ANOVA was conducted on the severity ratings of the problem instances. There was no main effect for protocol (F = 0.62, df = 1,14, n.s.) or user group (F = 0.00, df = 1,14, n.s.) and no interaction between protocol and user group (F = 0.09, df = 1,14, n.s.).

To investigate whether the two protocols identified the same distinct problems and what percentage of problems was identified by each protocol, the distribution of distinct problems identified by each method and by both methods was calculated for blind and sighted participants separately. Figure 1 shows that for all participants 27 % of the distinct problems were found by both CVP and RVP, with a slightly lower figure for sighted participants (23 %) than for blind participants (31 %). In total, RVP identified around 76 % of the distinct problems, whereas CVP only identified 51 % of the distinct problems.

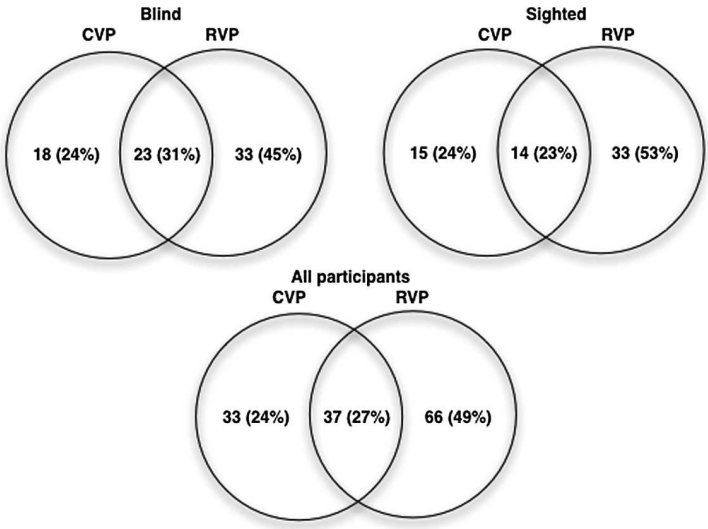

Fig. 1. Numbers and percentages of distinct problems identified for each protocol for the two user groups and for all participants across the four websites

The severity ratings of the problems identified by one protocol only and by both protocols were also investigated. The mean severity ratings are shown in Fig. 2, note that the mean severity ratings for all participants are the means for each user group weighted by the number of problems found by each user group. To investigate whether the problems by blind and sighted participants were rated more severely by one of the two protocols, the severity ratings of the problems that were found by both protocols were analysed. For blind participants, 23 problems were found by both protocols. The

mean severity of these problems when found using CVP was 2.43 (SD = 0.98), whereas when found using RVP it was 2.12 (SD = 0.65). A paired sample t-test showed that there was no significance difference between these ratings from the two protocols (t = 1.81, df = 22, n.s.). For sighted participants, 14 distinct problems were found by both protocols. The mean severity of these problems when found using CVP was 2.33 (SD = 0.93), whereas when found using RVP it was 2.40 (SD = 0.55). Again, a paired sample t-test showed that there was no significance difference between the ratings from the two protocols (t = -0.23, df = 13, n.s).

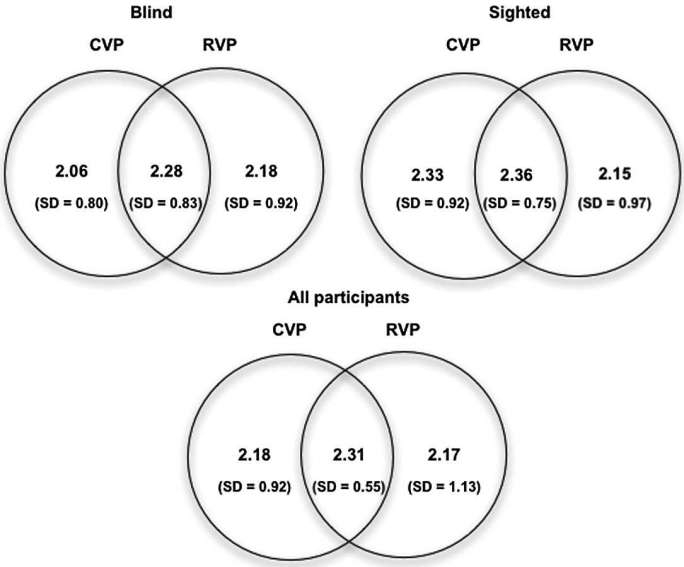

Fig. 2. Severity ratings of distinct problems identified for each protocol for the two user groups and for all participants across the four websites

To investigate the efficiency of the two protocols an analysis of the number of distinct problems identified per hour of evaluation time was conducted. A 2-way mixed ANOVA revealed that there was no main effect for protocol (F = 1.62, df = 1,14, n.s.). However, there was a main effect for user group (F = 30.17, df = 1,14, $p < 0.001$). The average number of distinct problems identified per hour for blind participants was 9.59 (SD = 4.36), whereas for sighted participants the average was 28.47 (SD = 9.96). Finally, there was no interaction between protocol and user group (F = 0.66, df = 1,14, n.s.).

To investigate the workload of undertaking the protocols for blind and sighted participants, an analysis of the NASA TLX scores was conducted. Table 1 shows the mean scores for each of the NASA TLX subscales and the overall mean score. A 3-way ANOVA (protocol x user group x NASA TLX subscale) revealed a significant main effect for protocol (F = 4.63, df = 1,14, $p < 0.05$). The overall mean NASA TLX score for CVP was 22.17 (SD = 13.76), whereas for RVP it was 22.77 (SD = 15.74). There

was no significant main effect for the NASA TLX subscale (F = 3.20, df = 1,14, n.s.) and user group (F = 2.28, df = 1,14, n.s.). Finally, there were no significant interactions between protocol, user group and NASA TLX subscales. To investigate whether there were any significant differences on any of the individual NASA TLX subscales between CVP and RVP, post hoc paired t-tests were conducted between each of the six pairs, but this failed to show any significance differences.

Table 1. Means on NASA TLX subscales for CVP and RVP

NASA TLX sub-scale	CVP Mean/SD	RVP Mean/SD
Mental Demand	40.50 (SD = 18.18)	40.44 (SD = 15.08)
Physical Demand	2.38 (SD = 5.44)	5.25 (SD = 10.68)
Temporal Demand	16.50 (SD = 12.25)	18.25 (SD = 20.13)
Performance	15.50 (SD = 8.83)	19.31 (SD = 16.12)
Effort	33.94 (SD = 18.29)	45.00 (SD = 20.43)
Frustration	24.19 (SD = 26.63)	38.38 (SD = 33.21)
Mean:	22.17 (SD = 13.76)	27.77 (SD = 15.74)

To investigate participants' attitudes towards the two protocols, an analysis of the ratings on the six questions answered after completing CVP was conducted. A 2-way ANOVA revealed that there was no main effect for question (F = 1. 38, df = 1,14, n.s.). There was a trend towards a significant difference for user group (F = 3.19, df = 1,14, p = 0.09). The average rating for questions asked about CVP from blind participants was 1.92 (SD = 0.96), whereas for sighted participants it was 2.46 (SD = 0.94), meaning sighted participants found CVP more disruptive than sighted participants. Finally, there was no interaction between questions and user group (F = 0.97, df = 1,14, n.s.).

Looking more specifically at the differences between the two user groups on the six questions (see the means in Tables 2 and 3), sighted participants found rating the severity of problems interrupted the flow of the task more than blind participants (Sighted mean: 3.0; Blind mean: 1.50) and also that it interrupted their concentration more (Sighted mean: 3.0; Blind mean: 2.00).

One-sample t-tests were conducted for each of the six questions for blind and sighted participants separately to investigate whether participants ratings were significantly above the "not at all" point and significant different from the midpoint of the scale ("moderately"). The one-sample t-tests that were compared with value 1 were one tailed, whereas the other one-sample t-tests were two tailed.

Table 2 shows the results of the one-sample t-tests for blind participants. It shows that blind participants found thinking out loud interrupted the flow of the task (Q1) and their concentration (Q3) significantly more than "not at all", but significantly less than "moderately". They found that rating problems for their severity interrupted their concentration significantly more than "not at all" but significantly less than "moderately" (Q4). Blind participants also found that performing the CVP was significantly different than the way they might do the tasks in real life (Q5). Further, they found that performing the CVP was significantly more tiring (Q6) than not performing it at all.

Table 2. One-sample t-tests for blind participants' questions about CVP

Question	Mean/SD	Test value = 1 df = 7 in all cases	Test value = 3 df = 7 in all cases
Protocol interrupt (Q1)	2.13 (SD = 0.64)	t = 4.97 p < 0.001	t = -3.86 p < 0.01
Rating interrupt (Q2)	1.50 (SD = 0.76)	t = 1.87 n.s.	t = -5.61 p < 0.01
Protocol concentration (Q3)	2.00 (SD = 1.07)	t = 2.65 p < 0.05	t = -2.65 p < 0.05
Rating concentration (Q4)	2.00 (SD = 0.93)	t = 3.06 p < 0.01	t = -3.06 p < 0.05
Protocol real life (Q5)	2.13 (SD = 1.64)	t = 1.94 p < 0.05	t = -1.51 n.s.
Protocol tiring (Q6)	1.75 (SD = 0.71)	t = 3.00 p < 0.01	t = -5.00 p < 0.01

Table 3 shows the results from the same one-sample t-tests for the sighted participants. It shows that sighted participants found that thinking aloud (Q1, Q3) and rating the problems for their severity (Q2, Q4) significantly interrupted the flow of the task and their concentration more than "not at all". They also found that performing CVP changed the way they perform the tasks compared with real life (Q5) and that it was significantly more tiring (Q6) than not performing it at all. In comparison to the moderate midpoint, the results showed that sighted participants found that thinking aloud interrupted the flow of the task (Q1) and their concentration (Q3), although the interruption was significantly less than the midpoint of the scale. Also they found performing CVP to be significantly less tiring (Q6) than the midpoint of the scale.

Table 3. One-sample t-tests for sighted participants' questions about CVP

Question	Mean/SD	Test value = 1 df = 7 in all cases	Test value = 3 df = 7 in all cases
Protocol interrupt (Q1)	2.25 (SD = 0.89)	t = 3.99 p < 0.01	t = -2.39 p < 0.05
Rating interrupt (Q2)	3.00 (SD = 0.93)	t = 6.11 p < 0.001	t = 0.00 n.s.
Protocol concentration (Q3)	2.25 (SD = 0.89)	t = 3.99 p < 0.01	t = -2.39 p < 0.05
Rating concentration (Q4)	3.00 (SD = 0.93)	t = 6.11 p < 0.001	t = 0.00 n.s.
Protocol real life (Q5)	2.50 (SD = 1.31)	t = 3.24 p < 0.01	t = -1.08 n.s.
Protocol tiring (Q6)	1.75 (SD = 0.71)	t = 3.00 p < 0.01	t = -5.00 p < 0.01

Participants were asked to rate how much thinking aloud during the tasks (for CVP) or during the replay of the task (during RVP) made them self-conscious about what they

were doing (on a scale from 1 = "Not at all" to 5 = "Very much"). A 2-way ANOVA revealed that there was no main effect for the protocol (F = 0.13, df = 1,14, n.s.) or for the user group (F = 0.09, df = 1,14, n.s.) and no interaction between protocol and user group (F = 2.02, df = 1,14, n.s.).

One-sample t-tests were also conducted for the self-conscious question comparing the participants' ratings for each protocol with a value of 1, (not making them self-conscious at all) and the midpoint value of 3 (making them moderately self-conscious). Table 4 shows the results from these one-sample t-tests. Blind participants found both protocols made them significantly more self-conscious about what they were doing than not doing them at all. However, when the results were compared with the midpoint value of 3, blind participants found that doing CVP made them significantly less self-conscious than the midpoint of the scale. Sighted participants found only that doing CVP made them significantly more self-conscious about what they were doing than not doing nothing at all.

Table 4. One sample t-test on ratings of self-consciousness of the two protocols, for blind and sighted participants

User group/ protocol	Mean/SD	Test value = 1 df = 7 in all cases	Test value = 3 df = 7 in all cases
Blind/CVP	1.87 (SD = 0.83)	t = 2.96, $p < 0.05$	t = -3.81, $p < 0.01$
Blind/RVP	2.25 (SD = 1.04)	t = 3.42, $p < 0.05$	t = -2.05, n.s.
Sighted/CVP	2.50 (SD = 1.07)	t = 3.97, $p < 0.05$	t = -1.32, n.s.
Sighted/RVP	1.88 (SD = 1.36)	t = 1.83, n.s.	t = -2.35, n.s.

Finally participants selected which of the two protocols they preferred undertaking. Five out of eight sighted participants preferred CVP and three preferred RVP, whereas of the eight blind participants four preferred CVP and four preferred RVP. A chi-square test showed that there was no difference between user groups in preference for the protocols and no difference overall in preference for one protocol over the other (X^2 = 0.25, df = 1, n.s.).

5 Discussion

This study investigated the use of two verbal protocols for conducting evaluations in terms of effectiveness, efficiency and the effects they had on blind and sighted participants.

In terms of effectiveness, the results indicate that RVP is more effective than CVP. RVP identified more distinct problems than CVP for both blind and sighted participants. In addition, there was no difference in the severity ratings of the distinct problems identified between the two protocols. Comparing the two protocols in terms of whether they identify the same problems, we found that only 27 % of the distinct problems were identified by both protocols. Van den Haak et al. [25–28] also compared overlap between the two protocols in their studies. The overlap in most of the studies

[25–27] was similar with the overlap reported in the study here, except for one study [28]. Unfortunately, van den Haak et al. did not specifically report the overlap between CVP and RVP, as they included other protocols in their studies. However, found that the overlap between protocols which included CVP and RVP ranged from 25 % to 39 %. In addition, in this study RVP revealed 76 % of the total number of distinct problems, whereas CVP revealed only 51 % of the total number of distinct problems, with very similar figures for both blind and sighted participants. Finally, there was no difference between the severity ratings of the distinct problems found by both protocols from either user group and the severity of the problems that RVP failed to uncover was relatively low.

Although CVP is the more commonly used protocol [8, 12, 19, 20, 22], in this study CVP only identified approximately half of the distinct problems, whereas RVP identified three quarters. This contradicts the results of previous studies conducted by van den Haak et al. [25–28], that compared the two verbal protocols with sighted participants and found that they were comparable in terms of effectiveness. One possible explanation as to why the results are different lies in what van den Haak et al. identify as a user problem. In their studies, van den Haak et al. relied on a combination of user identified problems (i.e. problems that users verbalized themselves as problems) and problems identified by experts from reviewing the videos after the evaluation with the participants. In this study we were more conservative in our definition of user problems, in that we only considered those that were verbalized by participants.

In terms of efficiency, there was no difference between the two protocols. However, there was a significant difference in efficiency between the two user groups. Sighted participants identified nearly three times the number of distinct problems per hour compared with blind participants. This is not surprising as blind users interact with websites differently from sighted users and typically take longer to complete tasks. In this study, the blind participants typically took three times as long to complete tasks as the sighted participants, results very much in line with results from the Disability Rights Commission investigation of web accessibility [5], and also in line with the difference in efficiency with sighted participants.

In terms of the effects of the protocols on participants, the NASA TLX showed that RVP demanded more workload than CVP for both blind and sighted participants. However were a number of differences between blind and sighted participants on their perceptions of the two protocols, with sighted participants finding the rating of the severity of problems more disruptive than blind participants. However, comparing the ratings of the blind and sighted participants separately against "not at all" disruptive and "moderately" disruptive points revealed that both groups did find that CVP interrupted the flow of the task and concentration somewhat.

Comments from blind participants on this disruption included:

"when I think aloud I may miss what JAWS is talking to me and I may forget what I was doing and where I was"

"when I was trying to find things I had to think aloud and interrupted my concentration … it is difficult and sometimes frustrating"

"I was not listening 100 % on JAWS … there is a lot of processing information I had to use a lot of senses"

"I was not listening 100 % on JAWS … there is a lot of processing information I had to use a lot of senses"

These comments highlight how blind participants found thinking aloud interrupted their concentration and may cause them miss output from the screen reader. It was difficult for them to think aloud while they were trying to process the output of the screen reader and perform the task at the same time.

Comments from sighted participants on the disruption included:

"… trying to think aloud did interrupt the flow of the task"
"…by verbalizing my thoughts through process I assumed I was missing something"

These comments highlight how sighted participants found that thinking aloud interrupted the flow of the task and their concentration.

The two protocols are comparable in terms of how self-conscious the participants were about what they were doing. There was no difference between user groups in preference for the protocols. Participants were also asked to explain their choice. Comments from participants who preferred RVP included:

"I found [RVP] more easy to follow during the replay of the task"
"it was easier to do the tasks [in RVP] in silence you were able to concentrate more on what you were doing … RVP was easier because it was easier to listen to VoiceOver"
"thinking aloud during the task was hard … forgetting what I was doing … it was a distraction … RVP was easier but demanded more time"

Comments from participants who preferred CVP included:

"It was my normal way … I talk to the screen regularly"
"because it's quicker"
"it's in real time … beneficial at the time"

The comments show that some participants found it easier to perform RVP, as it did not interrupt them, especially blind participants who had to process the output of the screen reader in addition to performing the protocol. However, other participants preferred CVP because it was quicker compared to RVP.

6 Conclusions

This study compared two verbal protocols, CVP and RVP, with blind and sighted participants. The two protocols were compared in terms of effectiveness, efficiency and the effect they have on participants. The study provides insight in terms of which verbal protocol is appropriate for use in studies with both blind and sighted participants.

The key results are that RVP outperforms CVP in terms of effectiveness but is no more efficient than CVP. RVP identifies more distinct problems and problem instances than CVP for both blind and sighted participants. Also, both of the protocols are comparable in terms of identifying more severe problems. Further, the study demonstrated that there was quite a low overlap in the problems between the two protocols

identified for both blind and sighted participants. In addition, RVP identified three-quarters of the total number of distinct problems, whereas CVP only identified half of the distinct problems. In terms of efficiency, the protocols are comparable.

Even though RVP created a significantly higher workload for participants and CVP was perceived as being somewhat disruptive of the flow of the task, there was no clear preference amongst participants for one protocol over the other, so these did not strongly differentiate between the protocols.

Our future research will examine whether there is difference into the type of problems that the two protocols reveal. Also, an investigation whether there is difference into the problems that the two user groups reveal will be conducted.

Acknowledgements. We thank the National Council for the Blind of Ireland for their assistance in running this study, and all the participants for their time. Andreas Savva thanks the Engineering and Physical Science Research Council of the UK and the Cyprus State Scholarship Foundation for his PhD funding.

Research Data Access. Researchers wishing to access the data used in this study should visit the following URL for more information:
http://www.cs.york.ac.uk/hci/as1517/

References

1. Boren, T., Ramey, J.: Thinking aloud: reconciling theory and practice. IEEE Trans. Prof. Commun. **43**, 261–278 (2000)
2. Borsci, S., Federici, S.: The partial concurrent thinking aloud: a new usability evaluation technique for blind users. Assistive technology from adapted equipment to inclusive environments—AAATE, vol. 25, pp. 421–425 (2009)
3. Bowers, V.A., Snyder, H.L.: Concurrent versus retrospective verbal protocol verbal for comparing window usability. In: Human Factors and Ergonomics Society Annual Meeting, pp. 1270–1274 (1990)
4. Chandrashekar, S., Stockman, T., Fels, D., Benedyk, R.: Using think aloud protocol with blind users: a case for inclusive usability evaluation methods. In: 8th international ACM SIGACCESS Conference on Computers and Accessibility, pp. 251–252. ACM, New York (2006)
5. Disability Rights Commission.: The Web: Access and inclusion for disabled people. The Stationery Office, London (2004)
6. Dumas, J.F., Redish, J.C.: A practical Guide to Usability Testing. Greenwood Publishing Group Inc., Westpost, CT, USA (1993)
7. Ericsson, K.A., Simon, H.A.: Protocol Analysis. MIT-press, Cambridge (1984)
8. Harrison, C., Petrie, H.: Severity of usability and accessibility problems in eCommerce and eGovernment websites. In: BryanKinns, N., Blandfor, A., Curzon, P., Nigay, L. (Eds.), People and Computers XX - Engage. pp. 255–262, Godalming: Springer-Verlag London Ltd (2007)
9. Krahmer, E., Ummelen, N.: Thinking about thinking aloud: a comparison of two verbal protocols for usability testing. IEEE Trans. Prof. Commun. **47**, 105–117 (2004)
10. Kuusela, H., Paul, P.: A comparison of concurrent and retrospective verbal protocol analysis. Am. J. Psychol. **113**(3), 387–404 (2000)

11. Lavery, D., Cockton, G., Atkinson, M.P.: Comparison of evaluation methods using structured usability problem reports. Behav. Inf. Technol. **16**, 246–266 (1997)
12. Lazar, J., Olalere, A., Wentz, B.: Investigating the accessibility and usability of job application web sites for blind users. J. Usability Stud. **7**, 68–87 (2012)
13. Lewis: Using the Thinking Aloud Method in Cognitive Interface Design, Technical report. IBM Research Center (1982)
14. NASA TLX: Task Load Index, http://humansystems.arc.nasa.gov/groups/tlx/
15. Nielsen, J.: Usability Engineering. Elsevier (1994)
16. Nørgaard, M., Hornbæk, K.: What do usability evaluators do in practice?: an explorative study of think-aloud testing. In: 6th Design Interactive systems conference, pp. 209–218, ACM, New York (2006)
17. Ohnemus, K.R., Biers, D.W.: Retrospective versus concurrent thinking-out-loud in usability testing. In: Human Factors and Ergonomics Society Annual Meeting, pp. 1127–1131 (1993)
18. Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S., Ashenfelter, K.T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 2381–2390. ACM, New York (2010)
19. Petrie, H., Kheir, O.: The relationship between accessibility and usability of websites. In: SIGHI conference on Human Factors in Computing Systems, pp. 397–406. ACM, New York (2007)
20. Power, C., Freire, A., Petrie, H., Swallow, D.: Guidelines are only half the story: accessibility problems encountered by blind users on the web. In: SIGHI Conference on Human Factors in Computing Systems, pp. 433–442, ACM, New York (2012)
21. Rogers, Y., Sharp, H., Preece, J.: Interaction Design: Beyond Human-Computer Interaction. John Wiley & Sons, Hoboken (2011)
22. Rømen, D., Svanæs, D.: Validating WCAG versions 1.0 and 2.0 through usability testing with disabled users. Univ. Access Inf. Soc. **11**, 375–385 (2012)
23. Shi, Q: A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. In: 5th Nordic conference on Human-Computer Interaction: Building Bridges, pp. 344–352. ACM, New York (2008)
24. Strain, P., Shaikh, A.D., Boardman, R.: Thinking but not seeing: think-aloud for non-sighted users. In: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, pp. 1851–1856. ACM, New York (2007)
25. van den Haak, M.J., De Jong, M.D., Schellens, P.J.: Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. Gov. Inf. Q. **16**, 1153–1170 (2004)
26. van den Haak, M.J., De Jong, M.D., Schellens, P.J.: Evaluating municipal websites: A methodological comparison of three think-aloud variants. Gov. Inf. Q. **26**, 193–202 (2009)
27. van den Haak, M.J., De Jong, M.D., Schellens, P.J.: Evaluation of an informational web site: three variants of the think-aloud method compared. Tech. Commun. **54**, 58–71 (2007)
28. van den Haak, M.J., De Jong, M.D., Schellens, P.J.: Retrospective vs concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav. Inf. Technol. **22**, 339–351 (2003)
29. van den Haak, M.J., De Jong, M.D.: Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. In: IEEE International Professional Communication Conference. pp. 285–287 (2003)
30. Wright, R.B., Converse, S.A.: Method bias and concurrent verbal protocol in software usability testing. In: Human Factors and Ergonomics Society Annual Meeting, pp. 1220–1224 (1992)