Proceedings Paper:
Savva, Andreas, Petrie, Helen orcid.org/0000-0002-0100-9846 and Power, Christopher Douglas orcid.org/0000-0001-9486-8043 (2016) Types of problems elicited by verbal protocols for blind and sighted participants. In: Proceedings of the 15th International Conference on Computers Helping People with Special Needs. Lecture Notes in Computer Science. Springer.

https://doi.org/10.1007/978-3-319-41264-1

To be published in Proceedings of the 15th International Conference on Computers Helping People with Special Needs, Linz Austria, July 13 – 15, 2006. This is a pre-publication version.

Types of Problems Elicited by Verbal Protocols for Blind and Sighted Participants

Andreas Savva, Helen Petrie, Christopher Power

Human-Computer Interaction Research Group, Department of Computer Science, University of York, YO10 5GH, UK
{as1517, helen.petrie, christopher.power}@york.ac.uk

Abstract.
Verbal protocols are often used in user-based studies of interactive technologies. This study investigated whether different types of problems are revealed by concurrent and retrospective verbal protocols (CVP and RVP) for blind and sighted participants. Eight blind and eight sighted participants undertook both CVP and RVP on four websites. Overall, interactivity problems were significantly more frequent in comparison to content or information architecture problems. In addition, RVP revealed significantly more interactivity problems than CVP for both user groups. Finally, blind participants encountered significantly more interactivity problems than sighted participants. The findings have implications for which protocol is appropriate, depending on the purpose of a particular study and the user groups involved.

Keywords. User-based studies • concurrent verbal protocol • retrospective verbal protocol • usability • accessibility • blind users

1 Introduction

User-based studies are regarded as the gold standard for assessing the usability and accessibility of interactive systems. Typically, users perform a verbal protocol while they undertake tasks with the system. The verbal protocol was first introduced in human computer interaction studies by Lewis [9], but its origins can be traced back to the work of Ericsson and Simon [4, 5] in cognitive psychology. Verbal protocols can offer insight into the users' thought processes, their problem solving strategies [10] and it can be an effective method for detecting the problems users encounter with an interactive system [7], [20]. Many usability textbooks have established the verbal protocol as a core component of usability testing practice [3], [10], [14].

A key aspect of the approach proposed by Ericsson and Simon [4, 5] is the passive role of the evaluator during the study. The only intervention by the evaluator is to remind participants to think aloud if they become silent. Nevertheless, some evaluators do not follow this approach and take a more active role [1], [11]. Boren and Ramey [1] proposed an approach to verbal protocols that is based on the speech-communication theory. The evaluator provides acknowledgement tokens such as "mm hm" or "uh-huh" to keep participants verbalizing their thoughts.

The verbal protocol can be performed either concurrently, concurrent verbal protocol (CVP), or retrospectively, retrospective verbal protocol (RVP). In CVP participants think out loud while doing the task, whereas in RVP participants first perform the tasks in silence and then they perform the verbal protocol, usually prompted by a video of themselves performing the tasks [10], [13]. Blind participants can also perform RVP by listening to an audio of their interaction with their screen reader.

Numerous studies have compared the two protocols in terms of the participants' task success or the number of problems revealed [2], [15, 16, 17, 18, 19]. However, there is little research into the differences in the types of problems that the two protocols reveal [16, 17, 18, 19] and these studies have been conducted only with sighted participants. A limitation of these studies is that a single website was used in each one and inconsistencies in the classification of usability problems that were used to categorize the problems across studies. In addition, some studies showed that RVP reveals more problems of a specific type but the results were not consistent across studies. Moreover, the results of these studies cannot be generalized to all people as the participant's ages in all studies were between 18 and 25.

Even though these studies provide a better understanding of the different problem types that the two protocols reveal, more comprehensive studies with a wider variety of websites need to be conducted. As far as blind participants are concerned, no work could be found comparing the two verbal protocols in terms of the problem types they reveal and how they differ from the problems sighted participants encounter. It is important to investigate the research methods assessing usability and accessibility of websites, as we can get more insights into which method can be considered a better option for studies with either blind or sighted users. This paper investigates whether there is difference in the problem types revealed by CVP and RVP and between blind and sighted users.

2 Method

2.1 Participants

Sixteen participants, eight blind and eight sighted, undertook the study. The two groups of participants were matched as closely as possible in terms of age, gender, operating system used, web experience and web expertise. The blind participants were six men and two women with a median age of 43 years (range 23 – 64); the sighted participants also comprised six men and two women with median age of 40 years (range 22 – 55). Five blind and five sighted participants were Windows users and three blind and three sighted participants were Mac OSX users. Participants rated their web experience using a five-point Likert item (1 = very low to 5 = very good). Blind participants' average rating was 4.0 (SD = 0.9), whereas for sighted participants it was 4.5 (SD = 0.5). Participants also rated their web expertise in the same way. Blind participants' average rating was 3.8 (SD = 0.9), whereas for sighted participants it was 3.6 (SD = 0.9).

All blind participants used screen readers to navigate the web. The five participants who used Windows used JAWS as their screen reader and the three participants who used Mac OSX used VoiceOver as their screen reader.

2.2 Websites and Tasks

Four websites from different domains were used in the study: a government website (www.gov.uk), a real estate website (www.rightmove.co.uk), an online shop (www.boots.com) and a news website (www.channel4.com). The websites included a range of different web design aspects such as headings, forms, tables, and links. The tasks included both navigation and data input. Each participant undertook one task on each website. The tasks 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 24kg.
- Channel4: Find which movie will be on Film4 at 9pm the day after tomorrow.

2.3 Procedure

The study was conducted 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). Participants were briefed about the study and then signed an informed consent form. Participants used their preferred operating system and browser in order to avoid any problems related to lack of familiarity with the technology. Blind participants also used their preferred screen reader and the appropriate version. With their permission, all the sessions were recorded using Morae 3.1 on Windows and ScreenFlow 4.0.3 on Mac OSX.

For each protocol, the researcher gave a standard demonstration of the protocol that the participant was about to perform. The participants then tried out the protocol for themselves on a practice website. The verbal protocol procedure was based on the Boren and Ramey [1] approach. For CVP, participants thought out loud as they performed the tasks. If they were quiet for more than 20 seconds, they were prompted with a general question such as "What are you thinking about?". However, there were cases when the prompts relied on the evaluator's discretion, particularly for the blind participants. There were occasions when participants were silent for extended period because they were listening to the screen reader. For example, participants were searching for a specific link in a list of links, which may have included more than one hundred links, thus the 20 second time interval would not appropriate on such an occasion. For RVP, participants first performed the tasks in silence, then reviewed them on the video (or audio for the blind participants) which was played back to them after the completion of each task.

Each time participants encountered what they considered to be a problem (be it with the website, the browser or the operating system), they were asked to describe it.

After completing both protocols, participants were asked to complete a demographic questionnaire and were debriefed about the study and any questions they had were answered.

The order of the tasks and the verbal protocols were counter-balanced within each user group, to minimize practice and fatigue effects.

2.4 Data Analysis

The video recordings were reviewed and problems were categorized using the classification of usability problems developed by Petrie and Power [12]. This involves four main types of problem: physical presentation, content, information architecture and interactivity. An additional type was added to deal with the problems encountered by blind participants, for problems involving incompatibilities between the browser and the screen reader, we named this category technology problems. We used the classification of problems by Petrie and Power [12], as it was more explicit but a similar categorization of problems to that used by van den Haak et al. [16, 17, 18, 19]. To distinguish the differences between the content, information architecture and interactivity, we considered interactivity problems those that break the interaction of the user with the website, information architecture those that are related with the organization and the structure of the information between and within the pages and content problems those that are associated with the information in the pages. Table 1 shows examples of each problem type from blind and sighted participants.

Table 1. Examples of each problem type from blind and sighted participants

	Blind Participants	Sighted Participants
Content	There is nothing about schools in the description of the house (P8)	The product description is limited. There is nothing about weight (P16)
Information Architecture	The structure of the movies is confusing. I cannot understand which of the two times is the correct one for the movie (P5)	The option to filter by schools is very deep in the site (P13)
Interactivity	The input of the maximum number of bedrooms does not have a label (P1)	The group weight options in the filtering are not very clear (P15)

Inter-coder reliability on the identification of problems was calculated on 10% of the video sessions. An additional evaluator, not involved in the study, independently extracted the problems from the videos. The reliability was calculated using the any-two agreement by Hertzum and Jacobsen [6]:

$$\frac{|Pi \cap Pj|}{|Pi \cup Pj|}$$

The any-two agreement is based on the number of problems the two evaluators have in common divided by the total number of problems they identified. P refers to number

of problems identified and i and j refers to the two evaluators. The conservative approach we followed in terms of the definition a problem resulted in 100% agreement on the identification of user problems.

Inter-coder reliability on the categorization of problems was calculated on 10% of the problems. Cohen's Kappa (K) [8] was calculated between one of the authors and a additional coder who was not involved in the study. Inter-coder reliability showed satisfactory levels of agreement for the categorisation of the problems with K = 0.883 for the main types of problems and K = 0.836 for the sub-type of problems.

For the main analysis of data, we compared only the problems that were encountered by both user groups. Thus, we included only the content, information architecture and interactivity problems, as blind participants did not encounter any physical presentation problems and sighted participants did not encounter any technology problems.

3 Results

A total of 260 instances of problems were reported across both protocols and all websites. To investigate whether there is difference between problem types that the two protocol reveal and whether there were differences between the problem types reported by the two user groups, an analysis of the instances of problems of each type was conducted. A 3-way ANOVA (verbal protocol x user group x type of problems) did not reveal any significant main effect for user group ($F = 3.19$, $df = 1,14$, n.s.). Thus, blind and sighted participants did not differ in the overall number of problems encountered. The analysis revealed a significant main effect for verbal protocol ($F = 5.30$, $df = 1, 14$, $p < 0.05$). The mean number of problem instances in CVP was 5.94 (SD = 2.02) per participant, whereas in RVP it was 8.50 (SD = 4.00). The analysis also revealed main effect of problem type ($F = 41.07$, $df = 1.46, 20.42$, $p < 0.001$, with Greenhouse-Geisser correction). Post-hoc comparison using t-tests with Bonferroni correction indicated that the mean number of interactivity problems (M = 9.06, SD = 4.43) per participant was significantly higher than the mean number of content problems (M = 2.50, SD = 2.00) and the mean number of information architecture problems (M = 2.88, SD = 1.75).

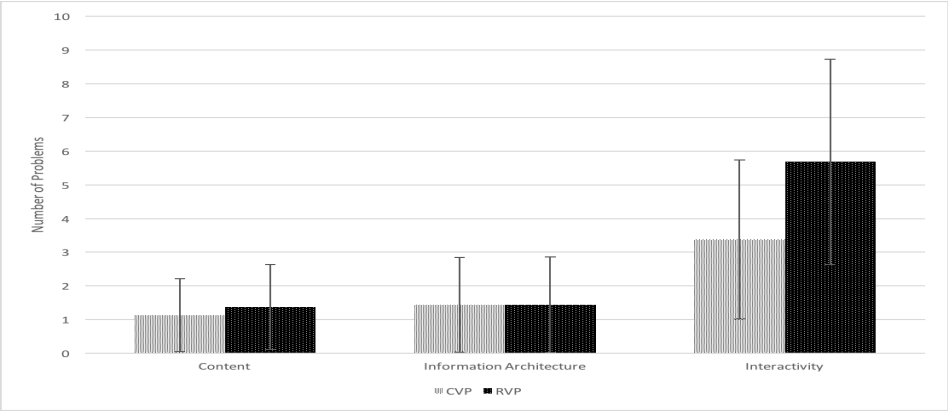

Fig. 1. Mean number of problems for the three problem types, for CVP and RVP

There was a significant interaction between verbal protocol and the problem type ($F = 4.29$, $df = 2, 28$, $p < 0.05$). Figure 1 shows the mean number of problems for the three problem types, for CVP and RVP, per participant. Post hoc paired sample t-tests showed there was there was a significant difference between protocols for the interactivity problems ($t = -2.79$, $df = 15$, $p < 0.05$). The mean number of interactivity problems identified using CVP was 3.38 (SD = 2.36), whereas in RVP it was 5.69 (SD = 3.05). None of the other comparisons were not significantly different.

There was also a significant interaction between user group and problem type ($F = 12.34$, $df = 1.46, 20.42$, $p < 0.001$, with Greenhouse-Geisser correction). Figure 2 shows the mean number of problems per problem type and user group. Post hoc sample-t-tests showed that there was a significant difference between blind and sighted participants on interactivity problems ($t = 3.47$, $df = 7$, $p < 0.05$). The mean number of interactivity problems encountered by blind participants was 12.00 (SD = 3.82), whereas for sighted participants it was 6.13 (SD = 2.42).

Further examination of the interactivity problems showed that there were interactivity problems that encountered only by blind participants and not by sighted participants. These problems included lack of feedback on user actions, labels missing on interactive elements, links that lead to external sites without warning, interactive elements not grouped clearly, lack of consistency between the interactive elements used, and input formats not clear. In addition, there were interactivity problems that were encountered more frequently by blind participants than by sighted participants. These included instructions on interactive elements not clear, options not complete, and elements not clearly identified as interactive or not.

There was no interaction between user group and verbal protocol ($F = 0.03$, $df = 1, 14$, n.s.). Finally, there was no significant three way interaction between problem type, verbal protocol and user group ($F = 1.13$, $df = 2, 28$, n.s.).

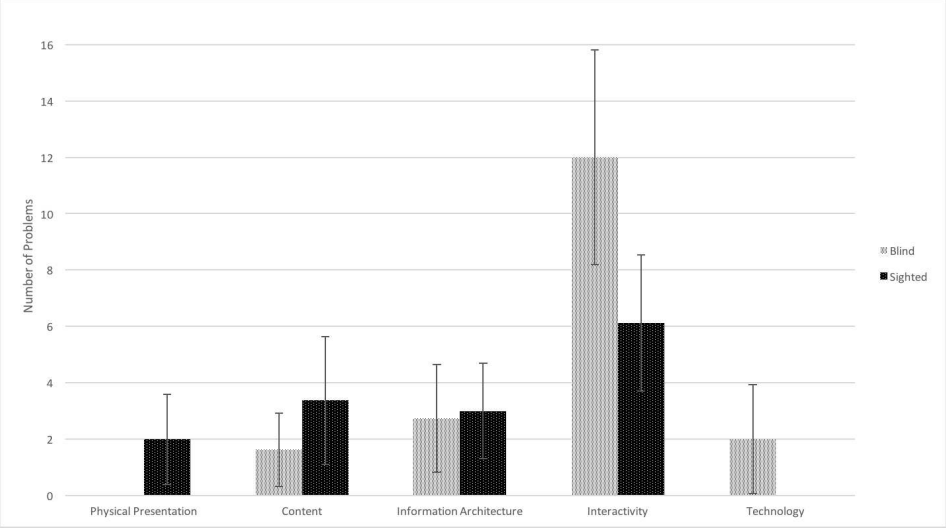

Fig. 2. Mean number of problems per problem type for blind and sighted participants

4 Discussion and Conclusions

This paper compared two verbal protocols, CVP and RVP, on whether they identify different types of problems. In addition, a comparison of the problem types revealed by blind and sighted users was conducted. The results indicate that RVP produced significantly more problems overall. There was also a significant difference between frequency of problem types. Interactivity problems were encountered significantly more often than content and information architecture problems. In addition, there was a significant interaction between protocol and the problem type: RVP revealed significantly more interactivity problems compared with CVP, with no differences in the other problem types. Finally there was a significant interaction between user group and problem type: blind participants significantly reported more interactivity problems than sighted participants, with no significant differences between the groups in the other problem types.

The difference in frequency in interactivity between blind and sighted participants comes from several sources. There were interactivity problems that only encountered by blind participants, for instance the lack of feedback on user actions and system progress, missing labels on interactive elements, and links that lead to external sites without warnings. There were also types of problems that were encountered by both user groups but which blind participants encountered more frequently than sighted participants. These included instructions on interactive elements not clear, and options not complete.

The study has provided a better understanding of the differences between the two verbal protocols in terms of the problem types the two protocols reveal. The results indicate that RVP may be considered a better option in user-based studies, particularly if the interest is in interactivity problems. However for studies interested in content or information architecture problems, either protocol is appropriate. We believe it is the first study to compare the type of problems found with the two protocols by blind and sighted participants and it has provided insights into the differences in terms of problem types between blind and sighted users.

Acknowledgements. We thank the National Council for the Blind of Ireland (NCBI) 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.

5 References

1. Boren, T., Ramey, J.: Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication 43(3), 261-278 (2000)
2. Bowers, V. A., Snyder, H. L.: Concurrent versus Retrospective Verbal Protocol for Comparing Window Usability. In: Human Factors and Ergonomics Society Annual Meeting, 34(17), 1270-1274 (1990)
3. Dumas J. S., Redish, J.: A practical guide to usability testing. Intellect Books (1999)

4. Ericsson, K. A., Simon, H.A.: Protocol analysis. Cambridge, MA: MIT Press (1993)
5. Ericsson, K. A., Simon, H.A.: Verbal reports as data. Psychological Review, 87(3), 215-253 (1980)
6. Hertzum, M., Jacobsen, N. E.: The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human Computer Interaction, 13(4), 421-443 (2001)
7. Jørgensen, A. H.: Thinking-aloud in user interface design: a method promoting cognitive ergonomics. Ergonomics 33(4), 501-507 (1990)
8. Landis, J. R., Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174 (1977)
9. Lewis, C.: Using the "thinking-aloud" method in cognitive interface design. IBM TJ Watson Research Center (1982)
10. Nielsen, J.: Usability engineering, Elsevier (1994)
11. Nørgaard, M., Hornbæk, K.: What Do Usability Evaluators Do in Practice?: An Explorative Study of Think-aloud Testing. In: 6th Conference on Designing Interactive Systems, 209-218 (2006)
12. Petrie, H., Power, C.: What Do Users Really Care About?: A Comparison of Usability Problems Found by Users and Experts on Highly Interactive Websites. In: SIGCHI Conference on Human Factors in Computing Systems, 2107-2116 (2012)
13. Preece, J., Sharp, H, Rogers, Y.: Interaction Design: beyond human-computer interaction. John Wiley & Sons (2015)
14. Rubin, J.: Handbook of Usability Testing, New York (1994)
15. Savva, A., Petrie, H., Power, C.: Comparing Concurrent and Retrospective Verbal Protocols for Blind and Sighted Users. In: Human-Computer Interaction-INTERACT, 55-71 (2015)
16. van den Haak, M. J, De Jong, M. D. T., Schellens, P. J.: Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. Interacting with Computers, 16(6), 1153-1170 (2004)
17. van den Haak, M. J., De Jong, M. D. T., Schellens, P. J.: Evaluation of an Informational Web Site: Three Variants of the Think-aloud Method Compared. Technical Communication, 54(1), 58-71 (2007)
18. van den Haak, M. J., De Jong, M. D. T., Schellens, P. J.: Evaluating municipal websites: A methodological comparison of three think-aloud variants. Government Information Quarterly, 26(1), 193-202 (2009)
19. van den Haak, M. J., De Jong, M. D. T., Schellens, P. J.: Retrospective vs concurrent think-aloud protocols: Testing the usability of an online library catalogue. Behavior & Information Technology, 22(5), 339-351 (2003)
20. Wright, P. C., Monk, A. F.: The use of think-aloud evaluation method in design. ACM SIGCHI Bulletin 23(1), 55-57 (1991)