Assistive Technology from Adapted Equipment to Inclusive Environments
P.L. Emiliani et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-042-1-421

The Partial Concurrent Thinking Aloud: A New Usability Evaluation Technique for Blind Users

Simone BORSCI [a, 1] and Stefano FEDERICI [b]
[a] *ECoNA - Interuniversity Centre for Research on Cognitive Processing in Natural and Artificial Systems, University of Rome 'La Sapienza', IT;*
[b] *Department of Human and Education Sciences, University of Perugia, IT;*

Abstract. The aim of this study is to build up a verbal protocol technique for samples of visual impaired users in order to overcome the limits of concurrent and retrospective protocols. Indeed, when blind users surf using a screen reader and talk about the way they interact with the computer, the evaluation is influenced by a structural interference. Users are force to think aloud and listen to the screen reader at the same time. The technique we improved, called Partial Concurrent Thinking Aloud (PCTA), integrates a modified set of concurrent verbalization and retrospective analysis. One group of 6 blind user and another group of 6 sighted users evaluated the usability of a website by PCTA. Estimating the number of users needed with an asymptotic test, we found out that the two groups had an equivalent ability of identifying usability problems, both over 80%. The result suggest that PCTA, even respecting the properties of classic verbal protocols, also allows to overcome the structural interference and the limits of concurrent and retrospective protocols when used with screen-reader users.

Keywords. Asymptotic test, Blindness, Thinking aloud, Usability evaluation.

Introduction

The diffusion of the universal design idea has required including users with disabilities in the usability evaluation process. This for two main reasons: First, since the accessibility is also a primary step in order to share information with disabled users and "open[s] up many opportunities for people with disabilities" [1], adapting internet technology to the users' need claims to improve the usability accordingly to disabled users' evaluations. Second, disabled users tend to have "unique and different computer interactions compared with their able-bodied counterparts" [2], opening up new issues for designers, usability practitioners, and researchers.

The researchers, moved by these new urgencies on the usability evaluation, have began to rethink some consolidated usability evaluation methods (UEMs), as the Thinking Aloud Protocol (TAP), adapting them to disabled users involved in the evaluations. In this study we purpose to split the TAP in two different experimental procedures of data collection: In the first one, the *concurrent* verbal protocol, data are collected during the decision task, instead of, in the second procedure, the *retrospective*

[1] Simone Borsci.

verbal protocol, after the decision task. These two kinds of verbal protocols are valid and reliable UEMs, even though just the retrospective thinking aloud method permits to overcome the structural interference when used with screen-reader users during the interaction with an interface [3,4].

There are a few comparative studies on concurrent and retrospective verbal protocols. The results of them show that there is not a significant difference between task performance and task completion time; but the retrospective TAP condition resulted in considerably fewer verbalizations in respect to concurrent verbalizations [5,6]. Even though these comparative studies have different points of view on verbal protocols, their attention is focused mostly on users' task performances and verbalizations, and on the TAP efficiency and efficacy in describing these two aspects. However, these studies do not consider the different cognitive processes activated by these two kinds of verbal protocols. Indeed, the concurrent thinking aloud protocol and the retrospective one are driven by different processes and categories of thought: The verbalization of the first one (*concurrent*) is focused on problems and strategies of a single surfing step; the verbalization of the other one (*retrospective*) is focused on descriptions influenced by user's experience on the entire evaluation process. Subjects use certain cognitive processes when they analyse and verbalize what they have done or why they have taken a certain decision 20 minutes before, and other processes when they verbalize while performing tasks, or just 5 seconds later. In the retrospective thinking aloud, with or without stimuli, using the long-term memory and making a cognitive reconstruction of their experience, users tell a story of their actions, strategies and problems. In the concurrent thinking aloud, users express their problems, strategies, stress, and impressions without the influence of a "rethinking" perception. In this sense, these two verbal protocols detect very different users' points of view: The retrospective TAP seems to be a more subjective measure (i.e. conscious mediated) than the concurrent one.

In general, in the usability evaluation it might be used both retrospective and concurrent TAP according to the study aims and goals. Nevertheless, when a usability evaluation is carried out with blind people, since using a screen reader and talking about the way of interacting with the computer implies a structural interference, several studies propose to use the retrospective TAP [2,7,8]. The use of retrospective TAP with disabled users remains only a functional solution, for two main reasons: First, it permits to overcome the user cognitive limitation, but it fails to analyse users' performance during an interaction, as the concurrent TAP does. Second, since the efficiency of concurrent technique greatly decreases when used with blind people in comparison to sighted users, practitioners prefer to use the retrospective model over the concurrent, even though, in this way, the number of verbalizations remarkably decreases.

Our hypothesis is that it is possible to reduce the screen reader influence without losing the advantages of the proximity within action, thinking and verbalization. In order to do so, we have used and improved a new TAP technique, called Partial Concurrent Thinking Aloud (PCTA), that unifies the advantages of both concurrent and retrospective models. Then, we will discuss PCTA properties, improve its setting, and we will estimate the number of users needed for a PCTA web usability evaluation with an asymptotic test.

1. The Asymptotic test

The estimation of a technique efficiency, or cost effectiveness, could be calculated with the well known Nielesen and Landauer [9] mathematical model. The Authors show that generally the least number of users required for usability evaluation techniques ranges from three to five: Adding users over this number does not provide an advantageous discovery of new problems in terms of costs-benefits. The Author estimate the number of users needed with the following formula:

$$Found_{(i)} = N(1-(1-\lambda)^{i}) \quad (1)$$

In (1), N is the total number of problems in the interface, λ is the probability of finding the average usability problem when running a single average subject test (i.e. individual detection rate), and *i* is the number of users. Some international studies [10-13] have shown that a sample size of 5 participants is sufficient to find approximately 80% of the usability problems in a system, when the individual detection rate (λ) is at least .30.

Using this mathematical model, it can be found the range of users required for a usability test and, therefore, it can be calculated the increase of problems found adding users to the evaluation. We applied this mathematical model to PCTA in order to estimate its efficiency, then, we compared the number of users needed for PCTA with the number needed for classic concurrent protocol evaluation. Finally, we estimated the PCTA efficiency both with blind and sighted users.

2. Properties and setting of the Partial Concurrent Thinking Aloud

Our aim is to build up a usability assessment technique eligible to maintain the advantages of concurrent and retrospective protocols while overcoming their limits. Therefore, we have analysed the PCTA technique's efficiency with both blind and sighted users. In order to do so, we composed the PCTA method into two sections, one concurrent and one retrospective. The first section is a modified concurrent protocol built up according to the three concurrent verbal protocols criteria described by Ericsson and Simon [8].

- First criterion: *Subjects should be talking about the task at hand, not about an unrelated issue.* In order to respect this rule, the time between problem retrieval, thinking and verbalization must be minimized to avoid the influence of a long perceptual reworking and the consequent verbalization of unrelated issues. Blind participants, using a screen reader, increase the time latency between identification and verbalization of a problem. To minimize this latency users are trained to ring a desk-bell that stops both time and navigation; during this suspension, users can verbalize their strategies and problems. This setting modification allows to avoid the cognitive limitation problem and the influence of perceptual reworking, also creating a "memory sign" for the retrospective analysis.
- Second criterion: *To be pertinent, verbalizations should be logically consistent with the verbalizations that just preceded them.* For any kind of user it is hard to be pertinent and consistent in a concurrent verbal protocol. Therefore, the practitioners could generally interrupt the navigation and ask for a clarification or stimulate the users to verbalize in a pertinent way. In order to do so and stop navigation to seeing impaired users, we propose to negotiate a specific physical sign with them: The practitioner, sitting behind the user, will put his hand on the user's shoulder. This physical sign grants the verbalization pertinence and consistence.
- Third criterion: *A subset of the information needed during the task performance should be remembered.* The concurrent model is based on the link between working memory and time latency. The proximity between the occurrence of a thought and its verbal report allows users to verbalize on the basis of their working memory.

The second PCTA section is a retrospective one in which users analyse those problems previously verbalized in a concurrent way. The memory signs, created by the users ringing the desk-bell, overcome the limits of classic retrospective analysis; indeed, these signs allow the users to be pertinent and consistent with their concurrent verbalization, thus avoiding the influence of long term memory and perceptual reworking. The PCTA's main disadvantage may consist in the fact that it interrupts "the natural task flow"; still we must consider that the main object of TAP evaluations consists in the verbalizations of problems, and not in the "natural flow" analysis. Even

classic TAP evaluations are affected by this same PCTA problem: The concurrent verbalization requested to users, in fact, is far to be "natural" to the interaction and it also tends to modify the "task flow." On the other hand, the retrospective model, since it is centred on the "natural task flow," is generally influenced by a strong perceptual reworking of problems and strategies.

3. Subjects and Procedures

As stated before, just 5 users are enough to find out about 80% of usability problems [9,12,16]. According to this criterion, we composed a sample divided into two groups: An experimental one with 6 blind participants and a control group with 6 sighted participants. All blind volunteers experienced in the same screen reader (JAWS).

Blind participants followed the same steps as the sighted, just with two differences: First, in order to guarantee the blind users' efficacy in the navigation, they were tested at home with their own technologies and their own screen reader (JAWS) settings. Second, the users in TAP analysis were trained to ring the desk-bell before the verbalizations of problems. The data were analysed comparing the kind of problems identified by the participants of both groups and estimating the PCTA efficiency between blind and sighted participants with the Neilsen and Landauer [9] mathematical model.

4. Results and conclusion

In order to improve the efficiency of PCTA with blind and sighted participants, we calculated the probability of finding the average usability problems running a single test (i.e., λ). For the experimental group λ was equal to .25, while for the control group was .27. Applying the formula (1) we estimated that using PCTA with the 6 users of each group we could find out over 80% of total problems: 82% for the experimental group of blind participants, and 84% for the control group of sighted participants. (Although sighted users have got a slightly higher ability of identifying problems (84%) than the blind ones (82%), such difference is negligible). We calculated that with a group of 15 participants we could have reached the 99% of usability problems for the control group and 98% for experimental one. Obviously, in this way we would have increased significantly the analysis costs in order to discover less than 20% more of usability problems. The proximity of λ value obtained by both the two groups (.25 for experimental and .27 for control group) to the average TAP λ value (.30), estimated by Nielsen with experimental studies involving large samples of users, provides evidence that PCTA guarantees the same efficiency properties of the classic thinking aloud. Moreover, the PCTA is a useful technique to assess usability with blind users, because it overcomes the structural interference imposed by the classic TAP that forces the user to concurrently think aloud and listen to the screen reader; at the same time, the PCTA also allows to avoid the influence of long term memory and perception unavoidable in the retrospective thinking aloud technique. PCTA seems to have a good efficiency with at least 6 users in both groups, rather than only 5 as Nielsen pointed out. Finally, both the experimental and the control groups seem to respect the tendency of data showed in international studies on the classical verbal protocols [12,16-18]. Even though the present study is based only on a summative evaluation (i.e. the

analysis of already published websites) and not on a formative one (i.e. the analysis of an interface during the user centered design process), our results still show that PCTA could be used in the usability evaluation with mixed samples of users, allowing disabled people, and in particular blind users, a partial concurrent analysis. We showed that, using PCTA, blind users' verbalizations of problems could be more pertinent and comparable to those given by sighted people who use a concurrent protocol. The use of PCTA could be widened to both summative and formative usability evaluations with mixed panels of users, thus extending the number of problem verbalizations according to disabled users' divergent navigation processes and problem solving strategies.

References

[1] K.P. Coyne and J. Nielsen, *Beyond ALT Text: Making the web easy to use for users with disabilities*, Nielsen/Norman Group Reports, 2001.

[2] S. Chandrashekar, D. Fels, T. Stockman and R.Benedyk, Using Think Aloud Protocol with Blind Users: A Case for Inclusive Usability Evaluation Methods. *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility* New York, NY: ACM, 2006

[3] Z. Guan, S. Lee, E. Cuddihy and J. Ramey, The validity of the stimulated retrospective think-aloud method as measured by eye tracking, SIGCHI, Conference on Human Factors in computing systems, 2006, Montréal, Canada,1253 – 1262

[4] M.J. Van den Haak, and M.D.T. De Jong, Exploring Two Methods of Usability Testing: Concurrent versus Retrospective Think-Aloud Protocols, IPCC 2003, International Professional Communication Conference Proceedings, 21 September, 2003, New York, USA.

[5] J.M. Hoc and J. Leplat, Evaluation of different modalities of verbalization in a sorting task, *International Journal of Man-Machine Studies*, **18** (1983), 283-306.

[6] V.A. Bowers and H.L. Snyder, Concurrent versus retrospective verbal protocols for comparing window usability, HFES, Human Factors Society 34th Meeting, 8-12 October, 1990, Santa Monica, USA, 1270-1274.

[7] P. Strain, A.D. Shaikh, and R. Boardman, Thinking but not seeing: think-aloud for non-sighted users, CHI '07, Human factors in computing systems, 2007, California, USA, 1851-1856.

[8] H. Takagi, S. Saito, K. Fukuda and C. Asakawa, Analysis of Navigability of Web Applications for Improving Blind Usability, *Comput.-Hum. Interact.*, **14** (2007), 13-37

[9] J. Nielsen and T.K. Landauer, A mathematical model of the finding of usability problems, Interact and CHI'93, conference on Human factors in computing systems,1993, Amsterdam, Netherland, 206-213.

[10] J. Nielsen, Heuristic evaluation, in: *Usability inspection methods*, J. Nielsen and R.L. Mack, eds, New York:Wiley, , 1994, pp. 25–62

[11] R.A. Virzi,(1990). Streamlining the design process: Running fewer subjects, Human Factors and Ergonomics Society 34th Annual Meeting, 1990, Santa Monica,USA, 291–294

[12] R.A. Virzi, Refining the test phase of usability evaluation: How many subjects is enough?, *Human Factors*, **34** (1992), 457-468.

[13] P. Wright and A. Monk, A cost-effective evaluation method for use by designers, *International Journal of Man–Machine Studies*, **35** (1991), 891–912.

[14] K.A. Ericsson, and H.A. Simon, Verbal reports as data, *Psychological Review* **87**(1980), 215 - 251.

[15] K.A. Ericsson, and H.A. Simon, *Protocol analysis: Verbal reports as data*, Cambridge, MA: MIT Press, 1993.

[16] J. Nielsen, Estimating the number of subjects needed for a thinking aloud Test, *International Journal of Human-Computer Studies*, **41** (1994), 385–397.

[17] J. Nielsen, Finding usability problems through heuristic evaluation, CHI'92, Conference on Human Factors in Computing Systems, 1992, California, USA, 373–380.

[18] C.W. Turner, J.R. Lewis, and J. Nielsen, Determining Usability Test Sample Size, in: *International Encyclopedia of Ergonomics and Human Factors*, (2 ed., Vol.3), W. Karwowski, ed, Boca Raton, FL: CRC Press., 2006, pp. 3084-3088.