Cogn Process (2010) 11:263–272
DOI 10.1007/s10339-009-0347-y

RESEARCH REPORT

Web usability evaluation with screen reader users: implementation of the partial concurrent thinking aloud technique

Federici Stefano · Simone Borsci · Gianluca Stamerra

Received: 17 May 2009 / Accepted: 28 October 2009 / Published online: 15 November 2009

Abstract A verbal protocol technique, adopted for a web usability evaluation, requires that the users are able to perform a double task: surfing and talking. Nevertheless, when blind users surf by using a screen reader and talk about the way they interact with the computer, the evaluation is influenced by a structural interference: users are forced to think aloud and listen to the screen reader at the same time. 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. 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 users and another group of 6 sighted users evaluated the usability of a website using PCTA. By estimating the number of necessary users by the means of an asymptotic test, it was found out that the two groups had an equivalent ability of identifying usability problems, both over 80%. The result suggests that PCTA, while 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. In this way, PCTA reduces the efficiency difference of usability evaluation between blind and sighted users.

Keywords Asymptotic test · Human computer interaction · Thinking aloud · Usability evaluation

Introduction

The spreading of the universal design idea has required users with disabilities to be include in the usability evaluation process. This for two main reasons: first, since the accessibility is a primary step in order to share information with disabled users and since it "opens up many opportunities for people with disabilities" (Coyne and Nielsen 2001), by adapting internet technology to the users' needs means 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" (Chandrashekar et al. 2006), opening up new issues for designers, for usability practitioners, and for researchers.

The researchers, pushed by this new approach on the usability evaluation, began to rethink some consolidated usability evaluation methods (UEMs), as the thinking aloud protocol (TAP), and started to adapt these techniques to the disabled users involved in the evaluations.

In the human computer interaction's (HCI) field, TAP, known as verbal protocol analysis, had a large application in the study of consumer and judgment making processes (Bettman 1979; Bettman and Park 1980; Biehal and Chakravarti 1982a, b, 1986, 1989; Green 1995; Kuusela et al. 1998).

Electronic supplementary material The online version of this article (doi:10.1007/s10339-009-0347-y) contains supplementary material, which is available to authorized users.

F. Stefano
Department of Human and Education Sciences, University of Perugia, Perugia, Italy

S. Borsci · G. Stamerra
ECoNA, Interuniversity Centre For Research on Cognitive Processing in Natural and Artificial Systems, University of Rome 'La Sapienza', Via dei Marsi, 00186 Rome, Italy

In describing this users-based evaluation processes, Hannu and Pallab (2000) state: "The premise of this procedure is that the way subjects search for information, evaluate alternatives, and choose the best option can be registered through their verbalization and later be analysed to discover their decision processes and patterns. Protocol data can provide useful information about cue stimuli, product associations, and the terminology used by consumers." Accordingly to this, we have split up the TAP in two different experimental procedures: the first one is the concurrent verbal protocol, collected during the decision task; the second procedure is the retrospective verbal protocol gathered after the decision task (Hannu and Pallab 2000).

By analysing the concurrent verbal protocol, Ericsson and Simon (1980) show that "verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they are obtained, are a valuable and thoroughly reliable source of information about cognitive processes". In this sense, the cognitive processes that generate verbalizations are a subset of the cognitive processes that generate the behaviour or the action. Ericsson and Simon (1993) have also identified three criteria that concurrent verbal protocol needs to satisfy:

"(1) Subjects should be talking about the task at hand, not about an unrelated issue. (2) To be pertinent, verbalizations should be logically consistent with the verbalizations that just preceded them. (3) A subset of the information heeded during the task performance should be remembered."

Guan et al. (2006) in their analysis have identified three main limits to the concurrent model: first, the act of speaking concurrently to the action may have a negative effect on the user's task performance. Second, the effort that the user makes to verbalize information while performing tasks might distract the subject attention and concentration. Third, the effort to fully verbalize the steps of the work might change the way that the user attends to the task components.

On the other hand, the retrospective thinking aloud collects the verbalization of a user's performance after the performance is over. The verbalization could take place without stimuli, which is likely to have a negative effect on the exhaustiveness of the comments produced, or with stimuli, i.e., supported by a recording of the performance (Guan et al. 2006; Van den Haak and De Jong 2003). In the stimuli-condition, after performing a web navigation silently, the users are asked to watch the recorded video of their performance and to verbalize the problems occurred during the interaction. Differently, in the without-stimuli-condition the users are asked to verbalize the problems occurred during the interaction without a support of the recording of their previous performance.

The introduction of the retrospective thinking aloud allows to overlap some of the limits of the concurrent protocol, even if it does not take into account the most specific property of concurrent model, i.e., the verbalization of thoughts based on working and short-term memory without the influence of long-term memory process and perception (Johnstone et al. 2006). In order to better understand these issues, we have to analyse both the comparative studies about verbal protocols and their applications with disabled people that we are going to discuss in the two following sub-sections.

Comparison of verbal protocol techniques

There are a few comparative studies on concurrent and retrospective verbal protocols (Hannu and Pallab 2000). The differences in findings of these studies are due to different measures adopted by the researchers to assess and compare the two kinds of TAP. Indeed, in order to compare different TAPs, researchers can consider a large number of factors, such as: the number of problems found (Hoc and Leplat 1983), the time of the users' performances (Bowers and Snyder 1990), the pertinence of users' verbalizations (Ericsson and Simon 1993; Van den Haak and De Jong 2003), the users' workloads (Van den Haak and De Jong 2003), and the degree of reactivity (i.e., when using the thinking aloud protocol, the reactivity of the participant might be different from usual—e.g., such a phenomenon occurs when subjects alter their performance due to their awareness of being observed by the technique adopted by the researcher).

Albeit the debate about the validity of concurrent and retrospective thinking aloud is still going on (see Guan et al. 2006 for the retrospective technique validity), in general, researchers claim that there is not a significant difference between task performance and task completion time; therefore, concurrent protocol analysis is usually preferred in usability evaluations, since it outperforms the retrospective one (see Hannu and Pallab 2000 for a comparison of methods). The retrospective TAP condition resulted in considerably fewer verbalizations in respect to the concurrent ones (Bowers and Snyder 1990; Hoc and Leplat 1983). Moreover, Van den Haak and De Jong (2003), comparing different users' task performances, show that participants in the concurrent TAP condition perform less successfully than the participants who work silently and verbalize in retrospect; the difference between the two TAPs is considerable both in terms of numbers of observable problems per participant and in the overall success rate for the tasks. On the other hand, as Ericsson and Simon (1993) show, retrospective data are less accurate than concurrent ones, and users' verbalizations in

retrospective condition are more focused on explanations and less on procedures, therefore resulting less pertinent than the concurrent verbalizations (Bowers and Snyder 1990). Hannu and Pallab (2000), comparing the effectiveness of concurrent and retrospective TAPs, show that concurrent analysis provides "a more insights into decision-making steps occurring between stimulus introduction and the final choice outcome" even though more statements about the final choice are provided in retrospective TAP (p. 387). Finally, Van den Haak and De Jong (2003) hypothesis is that the performance difference noticed between the two TAPs is mostly due to the different degree of reactivity and workload needed to the participants.

According to our opinion, the solution of the debate about validity and reliability of the TAPs cannot be found neglecting the different cognitive processes involved in the concurrent vs. retrospective technique. 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 the 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 min before, and other processes when they verbalize while performing tasks, or just 5 s later. In the retrospective thinking aloud, with or without stimuli, by 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 or frame-based represented (Minsky 1975)—than the concurrent one.

The think-aloud with screen reader users

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.

In general, in the usability evaluation both retrospective and concurrent TAP could be used according to the study aims and goals. Nevertheless, when a usability evaluation is carried out with blind people several studies propose to use the retrospective TAP: indeed, using a screen reader (an assistive technology software that attempts to convert text displayed on the screen in speech, sound icons, or a Braille output) and talking about the way of interacting with the computer implies a structural interference between action and verbalization (Guan et al. 2006; Strain et al. 2007; Takagi et al. 2007). Indeed, as Strain et al. (2007) have noticed, the use of a screen reader "leads to a significant challenge for the moderator, since the screen reader audio interferes with any dialogue between moderator and participant. Perceptual studies have shown that it is possible for humans to deal with two voices at once (the so-called "cocktail party effect"); however, due to cognitive limitations people often have a difficult time talking and listening at the same time" (p. 1853). These authors are referring to the Kemper, the Herman, and the Lian's study about the costs of doing two things at once for adults (2003).

Undoubtedly, basic cognitive studies provided a lot of evidence supporting the idea that individuals can listen, verbalize, or manipulate, and rescue information in multiple task condition. As Cherry (1953) showed, subjects, when listening to two different messages from a single loudspeaker, can separate sounds from background noise, recognize the gender of the speaker, the direction, and the pitch (cocktail party effect). At the same time, subjects that must verbalize the content of a message (attended message) listening to two different message simultaneously (attended and unattended message) have a reduce ability to report the content of the attended massage, while they are unable to report the content of the unattended message. Moreover, Ericsson and Kintsch (1995) showed that, in a multiple task condition, subjects' ability of rescuing information is not compromised by an interruption of the action flow (as it happens in the concurrent thinking aloud technique) thanks to the "Long Term Working Memory mechanism" of information retrieval.

Even if users can listen, recognize, and verbalize multiple messages in a multiple task condition and they can stop and restart actions without losing any information, others cognitive studies (Kemper et al. 2003) underlined that the overlap of activities in a multiple task condition have an effect on the goal achievement. Kemper et al. (2003), analysing the users abilities to verbalize actions in a multiple task condition, showed that the fluency of a user's conversation is influenced by the overlap of actions. Adults are likely to continue to talk as they navigate in a complex physical environment. However, the fluency of their conversation is likely to change: Older adults are likely to speak more slowly than they would if resting; young adults continue to speak just as rapidly when walking as when resting, but they adopt a further set of

speech accommodations, reducing sentence length, grammatical complexity, and propositional density. Just by reducing length, complexity, and propositional density adults free up working memory resources (*ivi*, p. 189).

We do not know how and how much the content of verbalizations could be influenced by the strategy of verbalization (i.e., the modification of fluency and the complexity in a multiple task condition). Anyway, we well know that users in the concurrent thinking aloud verbalize the problems in a more accurate and pertinent way (i.e., more focused on the problems directly perceived during the interaction) then in the retrospective one (Bowers and Snyder 1990; Ericsson and Simon 1993; Hoc and Leplat 1983; Van den Haak and De Jong 2003). The pertinence is granted to the user by the proximity of action-verbalization-next action; this multiple task proximity compels the subject to apply a strategy of verbalization that reduces the overload of the working memory. However, for blind users, this time proximity between action and verbalization is lost: the use of the screen reader, in fact, increase the time for verbalization (i.e., in order to verbalize, blind users must first stop the screen reader and then restart it).

Strain et al. (2007), in order to overcome the problems due to the screen reader use, suggested three different TAP methodologies with visual impaired users:

1. Traditional Retrospective Think-Aloud.
2. Modified Stimulated Retrospective Think-Aloud: The participant interacts with the interface without interruption. After attempting or completing the task, the moderator would ask the participant to slowly walk through the interface, and explain what he/she felt. During the walkthrough, the moderator could pause the screen reader as needed to probe for additional information. This technique was frequently used when testing prototypes.
3. Synchronized Concurrent Think-Aloud: The participant could choose to pause the screen reader audio in the middle of an interaction. The participant then discussed what was happening on the page and what they were experiencing. This method resulted in no conflicts with the screen reader audio since it was paused when dialogue was occurring. However, the natural task flow was interrupted. Synchronized method was preferred by participants who were comfortable thinking aloud and who were confident in stopping and starting the screen reader.

The use of retrospective TAP (and also of the Modified Stimulated Retrospective Think-Aloud) with disabled users remains only a functional solution, for two main reasons: first, it permits to overcome the user's cognitive limitations, but it fails to analyse the user's 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.

The Synchronized Concurrent Think-Aloud technique is a good solution, because it is focused on verbalization. This technique has been developed in order to avoid the screen reader interference and grant possibilities of verbalization to screen reader users. Nevertheless, the lack of a time limit for the user's verbalization allows avoiding the multiple task condition that is typical of concurrent processes; for this reason, in our opinion, the user's verbalization in Synchronized Concurrent condition is more similar to the retrospective than the concurrent one. Therefore, we expect that Synchronized technique, as the retrospective one, will provide a less effectiveness of data and a less pertinent users' verbalizations (Bowers and Snyder 1990; Ericsson and Simon 1993; Hannu and Pallab 2000).

Our hypothesis is that it is possible to reduce the screen reader influence (structural interference) without losing the advantages of the proximity within action, thinking, and identification of the problems (pertinence of users' verbalization). In order to do so, we have used and improved a new TAP technique, called Partial concurrent thinking aloud (Borsci and Federici 2009), that unifies the advantages of both concurrent and retrospective models. Then, we will discuss PCTA properties, improve its setting, and will estimate the number of users needed for a PCTA web usability evaluation with an asymptotic test.

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 (see the Fig. 1).

The first section is a modified concurrent protocol built up according to the three concurrent verbal protocols criteria described by Ericsson and Simon (1993).

– The first criterion is: *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

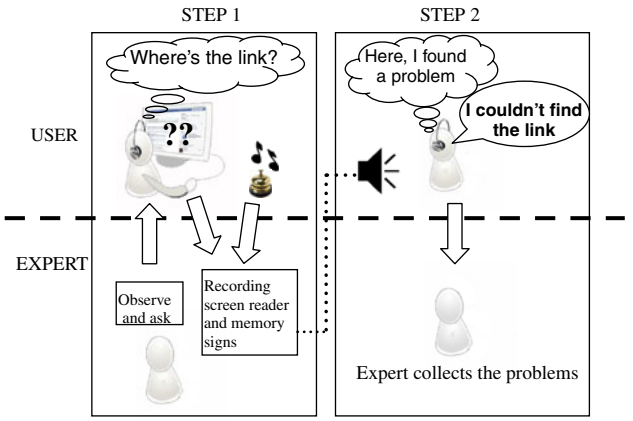

Fig. 1 Evaluation process of the partial concurrent thinking aloud. The PCTA technique is composed of two sections, one concurrent and one retrospective with different characteristics. In the first step [see below the frame "STEP 1"], the screen reader users, during a website interaction, create a memory sign (i.e., ringing a desk bell) each time they find a problem. In this concurrent step, the memory signs and the screen reader actions are audio–video recorded. In the second step [see below the frame "STEP 2"], the memory signs and the recorded stimuli facilitate users to recall those problems they previously identified. In this second step, users are involved in a retrospective analysis where they are invited to verbalize the problems they found

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 create a memory sign (i.e., ring the bell) and restart immediately the navigation. This setting modification allows avoiding the cognitive limitation problem and the influence of perceptual reworking, also creating a memory sign for the retrospective analysis.

– The second criterion is: *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[1] 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 screen reader 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.
– The third criterion is: *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 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.

As it also happens for the Synchronized protocol (Strain et al. 2007), 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 verbalizing problems, and not in the "natural flow" analysis. Even classic TAP evaluations are affected by this same PCTA problem: the concurrent verbalizations requested to users, in fact, are far to be "natural" to the interaction and they also tend 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.

As stated before, we are proposing three steps for PTCA evaluation:

– First: in order to minimize proximity between action, thoughts and verbalization, visual impaired users interrupt the navigation ringing a desk-bell next to the mouse (i.e., memory sign).
– Second: practitioners can touch users' shoulders with a hand as a physical sign (negotiated during training) in order to interrupt the navigation and ask about the action performed.
– Third: the retrospective session analysis is focused on those memory signs created during the concurrent session analysis.

Methods

Participants

Eighteen volunteers were selected, from students of University of Rome "La Sapienza", as a sample group: 8 blind and 12 sighted users. All blind volunteers needed to be experienced in the JAWS screen reader, and they have to set Jaws in order to read all graphic elements.

This sample of volunteers was tested using Sect. 7 of the European Computer Driving License (http:\\www.ecdl.com , [ECDL]) test that evaluates users' web navigation skills in

[1] Even if any interruption of the natural task flow is avoided in the Thinking Aloud, the moderator can make questions to the user in order to obtain pertinent verbalization of the problems.

a score range from 0 to 36 points (even though some international studies use self-questionnaires in order to recruit information on users' skills. We choose ECDL test because it is a valid and reliable international instrument that guarantees an estimation of users' navigation skills). The sample mean obtained by the ECDL test is 24 points (SD = 3.39). Our goal is not the estimation of users' skill level per se, but the selection of blind and sighted users by the means of an ECDL test score one point under and one point over the mean of our sample.

International studies show that a sample of 5 users is enough to get an evaluation able to find out about 80% of usability problems (Virzi 1992; Nielsen and Landauer 1993; Nielsen 1994a). Adopting this criterion, we composed a final sample divided into two groups: an experimental one with 6 blind participants and a control group with 6 sighted participants.

Apparatus

The apparatus of the experimental setting was set up as follows:

- Target web site: www.carabinieri.it (see Appendix 1a of Electronic supplementary material);
- Training web site: www.serviziocivile.it
- Browser: Internet Explorer 6;
- Internet connection ADSL 4 MB;
- Computer: PC AMD Athlon 64 (3,200 MHz)
- Monitor: Philips 190S LCD 19";
- Screen reader: Jaws;
- Screen recorder: CamStodio 20;
- Audio: Two amplifiers;
- Audio recorder: Digital Zoom h2
- Digital Camera: Nikon L2;
- Time: Stopwatch
- Support tools: Desk bell.

Procedure

The control group

Each participant of the control group was tested in the Psychology & Cognitive Lab of the University of Rome "La Sapienza". Each user was involved in a 20-min training session, with an explanation of the study goals, and in a simulation of a TAP website evaluation with 5 scenarios. The http://www.serviziocivile.it was used as a training interface. Then, the participants started the evaluation of the target website: http://www.carabinieri.it. Five tasks were presented as the experimental scenario (see Appendix 1b of Electronic supplementary material). Both the training website (i.e., http://www.serviziocivile.it) and the target website (i.e., http://www.carabinieri.it) are declared accessible by the Italian National Center for Informatics in Public Administration (http://www.pubbliaccesso.gov.it/logo/elenco.php). Once the TAP analysis was over, participants were invited to watch their concurrent evaluation recording (by screen recorder and video camera) and to start with the retrospective analysis adding any needed verbalization.

The experimental group

Each participant followed the same steps as the control group participants, 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) set in order to read all the text and graphic elements. Second, the users in TAP analysis were trained to ring the desk-bell any time they would have found a problem: this tool was used in order to create the memory signs needed for the subsequent retrospective analysis. In the retrospective steps, users were invited to listen the screen reader and their memory signs recorded in the concurrent step (by audio recorder) in order to verbalize the problems.

The data were analysed by comparing the kind of problems identified by the participants of both groups and estimating the PCTA efficiency between blind and sighted participants with the Nielsen and Landauer (1993) mathematical model.

The expert analysis of problems severity

Five experts, with more than 5 years of experience in usability evaluation of websites, were involved in an independent analysis of the problems found by the two groups of users, in order to rate their severity. The rate ranges from 1 (minor problems) to 3 (high problem) following the indication of Sears' comparative study (1997), in which a quite similar scale of problems severity is used to compare different cognitive walkthrough techniques.

Data analysis

All the data were processed using SPSS 16.0 for Windows, as follows:

- *Descriptive analysis*—frequency analysis of the usability problems found by the two groups. Then, all problems were weighted according to the expert rating scale of the problem severity.
- *Spearman's correlation analysis*—the score obtained by each user in the ECDL test was correlated to the number of user's verbalizations.

Then, an *asymptotic test* based on the Nielsen and Landauer (1993) mathematical model, was carried out on the problems identified by the users in order to esteem the technique efficiency, or cost effectiveness. Nielsen and Landauer (1993) 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 estimates the number of users needed with the following formula:

$$\text{Found}_{(i)} = N(1 - (1 - \lambda)^i) \quad (1)$$

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 (Nielsen 1994b; Virzi 1990, 1992; Wright and Monk 1991) 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. As an example, if for a 5 users evaluation λ equals .30, applying the formula (1), practitioners can estimate whether these 5 users are enough for an efficient assessment or, otherwise, how many n users are needed to increase the percentage of usability problems, as follows:

$$\text{Found}_{(5)} = 1 - (1 - 0.3)^5 = .83 \quad (2)$$

The problems rate obtained in this example with 5 users is .83 (i.e., 83% of usability problems). Afterwards, it can be estimated the increase of problems detection rate adding more users to this sample of five, as reported in Fig. 2.

The analysis of this hypothetical sample shows that almost 100% of usability problems can be found with 15 users, considering that: with just 5 users the likelihood of problems discovery is equal to 83%, and in order to discover less than 20% more of usability problems, not yet identified, at least ten more users need to be added to the evaluation.

We applied this mathematical model to PCTA in order to estimate its efficiency, and then we compared the number of users needed for PCTA with the number needed for classic concurrent protocol evaluation. In the end, we estimated the PCTA efficiency both with blind and sighted users.

Results and discussion

Analysis of the problems found

The experimental group found out 31 usability problems in total, while the control group users only 26; the two groups shared 12 highlighted problems. In the control group, 16 usability problems were detected by only one participant (i.e., 61% of total problems found), in the experimental group one participant detected 22 problems (i.e., 70% of total problems found).

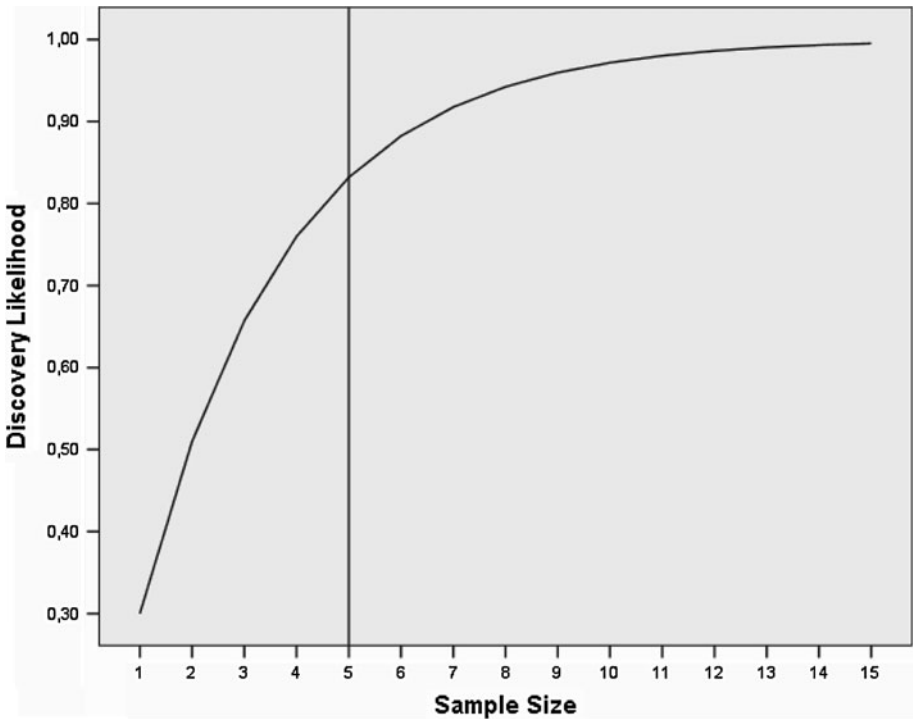

Fig. 2 Shows the asymptotic behaviour of discovery likelihood in relation to our hypothetical sample with $\lambda = .30$

The expert analysis of problems severity shows that 90% of problems found have a medium or high severity in line with the Nielsen (1992) and Virzi (1992) idea that users involved in TAP tend to find first the high-severity usability problems rather than the less relevant ones. Both the two groups found the same 3 minor problems. The 5 problems identified only by experimental group have an expert rate of severity equal to: medium for 3 problems and high for 2 problems. The difference between the number and the typology of problems found by the two groups seems to underline the importance of evaluations with disabled users, who tend to widen the number of problems found, thanks to a divergent process of navigation and different strategies of exploration, compared to users without disability (Chandrashekar et al. 2006).

Our interest is not to show that with PCTA screen reader users find more problems then sighted users, but that these problems have medium or high severity as in the classic concurrent technique and that screen reader users might find problems that sighted users did not find, enlarging the analysis.

It is interesting to note on a side that there is an inverse correlation ($P < .05$) between the score obtained by each blind user in Sect. 7 of the ECDL test and the number of his/her verbalized problems (Table 1).

In the experimental group, the participants with higher scores in Sect. 7 of the ECDL test (those with a greater expertise of navigation) verbalized a lower number of problems compared to the participants with a lower score in expertise of navigation, who, on the other hand, exceed in the verbalizations. This correlation between more and less expert users was not found in the control group. Such result nevertheless could be due to bias of the ECDL test with this type of disability.

Efficiency analysis

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. These results are expressed graphically by the Figs. 3 and 4 that show the proportion of usability problems found with increasing numbers of participants up to 15 users.

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

Table 1 Shows Pearson's correlation coefficient between problems verbalized by each screen reader user and the score obtained in Sect. 7 of ECDL test

Problems found by each blind user	ECDL test score of each blind user
Pearson correlation	−0.839[a]
Sig. (two-tailed)	0.037
N	6

[a] Correlation is significant at the 0.05 level (two-tailed)

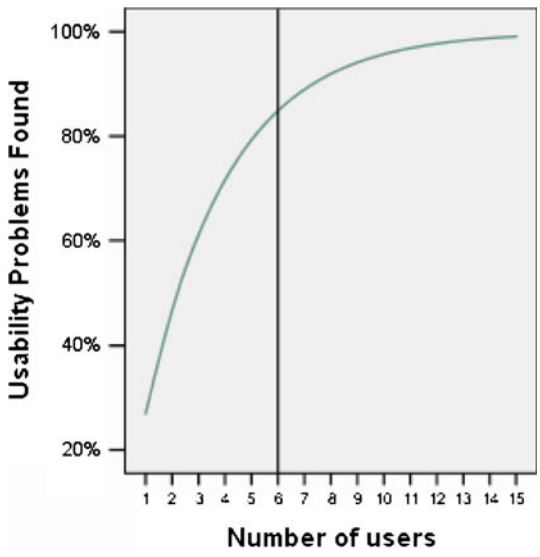

Fig. 3 Experimental group: proportion of usability problems found with increasing numbers of subjects ($\lambda = .25$) up to 15 users. The experimental group was formed by 6 participants who found the 82% of the usability problems

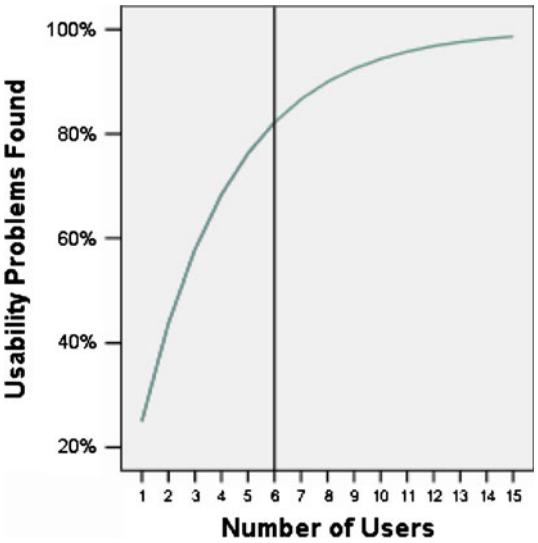

Fig. 4 Control group: proportion of usability problems found with increasing numbers of subjects ($\lambda = .27$) up to 15 users. Our control group was formed by 6 participants who found the 84% of the usability problems

experimental and the control groups seem to respect the tendency of data showed in international studies on the classical verbal protocols with 15 users (Nielsen 1994a; Turner et al. 2006; Virzi 1992).

Conclusion

The growing need in the HCI field for involving disabled users in the usability evaluation process has brought us to elaborate an integrated technique, the PCTA. This technique shows good analysis' properties and efficiency compared to the ones of classical verbal protocols.

Even though the present study is based only on a summative evaluation (i.e., the analysis of already published websites) rather than on a formative one (i.e., the analysis of an interface during the user centred 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. In the usability evaluation with blind people, the retrospective thinking aloud is often adopted as a functional solution to overcome the structural interference due to thinking aloud and hearing the screen reader imposed by the classic thinking aloud technique; such a solution has yet a relapse in the evaluation method because, as it stated before, the concurrent and the retrospective protocols measure usability from different points of view, one mediated by navigation experience (retrospective) one more direct and pertinent (concurrent). 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

Bettman JR (1979) An information processing theory of consumer choice. Addison-Wesley, Reading Cambridge

Bettman JR, Park CW (1980) Effects of prior knowledge and experience and phase of the choice processes on consumer decision processes: a protocol analysis. J Consum Res 7:234–248

Biehal G, Chakravarti D (1982a) Experiences with the Bettman-Park protocol coding scheme. J Consum Res 8:442–448

Biehal G, Chakravarti D (1982b) Information-presentation format and learning goals as determinants of consumers' memory retrieval and choice processes. J Consum Res 8:431–441

Biehal G, Chakravarti D (1986) Consumers' use of memory and external information in choice: macro and micro processing perspectives. J Consum Res 12:382–405

Biehal G, Chakravarti D (1989) The effects of concurrent verbalization on choice processing. J Mark Res 26:84–96

Borsci S, Federici S (2009) The partial concurrent thinking aloud: a new usability evaluation technique for blind users. In: Emiliani PL, Burzagli L, Como A, Gabbanini F, Salminen A-L (eds) Assistive technology from adapted equipment to inclusive environments—AAATE 2009. IOS Press, Amsterdam, pp 421–425

Bowers VA, Snyder HL (1990) Concurrent versus retrospective verbal protocols for comparing window usability. Human Factors Society 34th Meeting, 8–12 October 1990 HFES, Santa Monica, pp 1270–1274

Chandrashekar S, Fels D, Stockman T, Benedyk R (2006) 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. ACM, New York

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979

Coyne KP, Nielsen J (2001) Beyond ALT text: making the web easy to use for users with disabilities. Nielsen/Norman Group Reports

Ericsson KA, Kintsch W (1995) Long-term working memory. Psychol Rev 102:211–245

Ericsson KA, Simon HA (1980) Verbal reports as data. Psychol Rev 87:215–251

Ericsson KA, Simon HA (1993) Protocol analysis: verbal reports as data, Revised edn. MIT Press, Cambridge

Green A (1995) Protocol analysis. Psychologist 8:126–129

Guan Z, Lee S, Cuddihy E, Ramey J (2006) The validity of the stimulated retrospective think-aloud method as measured by eye tracking. Proceedings of the SIGCHI conference on human factors in computing systems, pp 1253–1262

Hannu K, Pallab P (2000) A comparison of concurrent and retrospective verbal protocol analysis. Am J Psychol 113:387–404

Hoc JM, Leplat J (1983) Evaluation of different modalities of verbalization in a sorting task. Int J Man-Mach Stud 18:283–306

Johnstone CJ, Bottsford-Miller NA, Thompson SJ (2006) Using the think aloud method (CognitiveLabs) to evaluate test design for

students with disabilities and English language learners. University of Minnesota, National Center on Educational Outcomes, Minneapolis

Kemper S, Herman RE, Lian CHT (2003) The costs of doing two things at once for young and older adults: talking while walking, finger tapping, and Ignoring Speech or Noise. Psychol Aging 18:181–192

Kuusela H, Spence MT, Kanto AJ (1998) Expertise effects on prechoice decision processes and final outcomes: a protocol analysis. Eur J Mark 32:559–576

Minsky M (1975) A framework for representing knowledge. In: Winston P (ed) The psychology of computer vision. McGraw-Hill, New York, pp 211–277

Nielsen J (1992) Finding usability problems through heuristic evaluation. In: CHI conference on human factors in computing systems. ACM, New York, pp 373–380

Nielsen J (1994a) Estimating the number of subjects needed for a thinking aloud test. Int J Hum-Comput Stud 41:385–397

Nielsen J (1994b) Heuristic evaluation. In: Nielsen J, Mack RL (eds) Usability inspection methods. Wiley, New York, pp 25–62

Nielsen J, Landauer TK (1993) A mathematical model of the finding of usability problems. In: Ashlund S, Mullet K, Henderson A, Hollnagel E, White E (eds) Proceedings of the InterCHI'93 conference. ACM, New York, pp 206–213

Strain P, Shaikh AD, Boardman R (2007) Thinking but not seeing: think-aloud for non-sighted users. CHI '07 extended abstracts on Human factors in computing systems, ACM, New York

Takagi H, Saito S, Fukuda K, Asakawa C (2007) Analysis of navigability of web applications for improving blind usability. Comput-Hum Interact 14:13–37

Turner CW, Lewis JR, Nielsen J (2006) Determining usability test sample size. In: Karwowski W (ed) International encyclopedia of ergonomics and human factors, vol 3, 2nd edn. CRC Press, Boca Raton, pp 3084–3088

Van den Haak MJ, De Jong MDT (2003) Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols, IEEE international professional communication conference proceedings, Piscataway

Virzi RA (1990) Streamlining the design process: running fewer subjects. In: Human factors and ergonomics society 34th annual meeting. Human Factors and Ergonomics Society, Santa Monica, pp 291–294

Virzi RA (1992) Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 34:457–468

Wright P, Monk A (1991) A cost-effective evaluation method for use by designers. Int J Man-Mach Stud 35:891–912