

Spring 6-20-2012

Concurrent and retrospective think-aloud protocols for information systems research

Ultan Sharkey

National University of Ireland Galway, Ireland, ultan.sharkey@nuigalway.ie

Thomas Acton

LERO @ National University of Ireland Galway, Ireland, thomas.acton@nuigalway.ie

Kieran Conboy

LERO @ National University of Ireland Galway, Ireland & UNSW, Australia, kieran.conboy@nuigalway.ie

Follow this and additional works at: <http://aisel.aisnet.org/bled2012>

Recommended Citation

Sharkey, Ultan; Acton, Thomas; and Conboy, Kieran, "Concurrent and retrospective think-aloud protocols for information systems research" (2012). *BLED 2012 Proceedings*. 31.

<http://aisel.aisnet.org/bled2012/31>

This material is brought to you by the BLED Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in BLED 2012 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Concurrent and retrospective think-aloud protocols for information systems research

Ultan Sharkey

National University of Ireland Galway, Ireland
ultan.sharkey@nuigalway.ie

Thomas Acton

LERO @ National University of Ireland Galway, Ireland
thomas.acton@nuigalway.ie

Kieran Conboy

LERO @ National University of Ireland Galway, Ireland
& UNSW, Australia
kieran.conboy@nuigalway.ie

Abstract

A resurgence of interest in qualitative approaches to research in online consumer behaviour studies has spawned a need for structured approaches in the use of appropriate research techniques. A mixed method approach to research questions can often lead to a greater clarity of answer as well as exposing the researcher to new 'tools of enquiry'. This paper describes a mixed method approach using qualitative techniques to pre-test software and task design for a field experiment and to pre-test a questionnaire for data collection. In the study, think-aloud protocol (TAP) sessions were adopted to test software. The study required participants to complete an online shopping task and thereafter complete a questionnaire. Twelve TAP sessions were carried out. The sessions enabled the researchers to verify both the propriety of the task and the face validity of the questionnaire. Here, the operation of the TAP methodology and a description of its use as an observational technique in testing an online questionnaire and online task-based research are presented

Keywords: think-aloud protocol, TAP, observational technique, cooperative evaluation, concurrent, retrospective, process tracing, talk-aloud, verbal protocols

1 Introduction

The study described here investigated the effects of product presentation modes on decision making behaviour and consumer experience in online shopping environments. A three-stage multi-method approach was taken. First, pilot experiments were conducted to gain an initial understanding of the online consumer shopping process in the particular context of the study. Subsequently, a verbal protocol method known as Think-Aloud Protocol (TAP) was conducted on software created for the study, which was then followed by implementation in a main experiment. The software was designed to allow participants complete a shopping decision task, and included a questionnaire for data collection. For the main experiment, the software was launched in conjunction with Ireland's largest online book retailer, requesting all visitors to the retailer's home page to participate in the study over a two week period. Those that chose to participate were funnelled into the online shop software created for the study.

Twelve TAP sessions were conducted during the second stage of this research. Each of the TAP sessions comprised a double application of the TAP method, first using the concurrent form of TAP followed by retrospective TAP. These sessions facilitated verification that the software allowed participants to complete the shopping task and questionnaire. It also provided the opportunity to test the face validity of the questionnaire and the task instructions. Further, early insights were gained into the behaviour of the TAP participants for the shopping decision scenario. As diagrammed in Figure 1-1 below the TAP sessions formed the intermediate stage between a series of earlier pilot experiments and before the later main experiment conducted in the field.

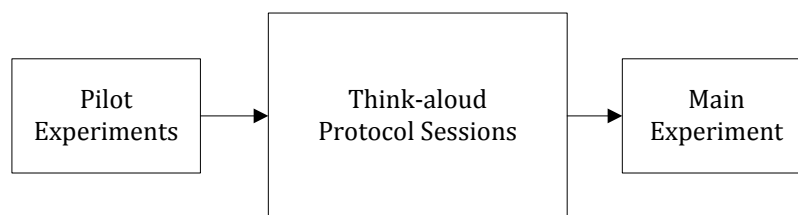


Figure 1-1 Chronological progression of the larger research project

This paper focuses on both the implementation and the value of concurrent and retrospective TAP sessions in a multi-stage multi-method approach to our study. Though the verbal protocol family of methods have had success in psychology and consumer behaviour studies, their use within information systems (IS) has been more limited. This paper describes the use of the think-aloud method in an IS scenario and provides specific support to IS researchers for using the think-aloud method.

2 Verbal Protocols

Developed from the idea of tracing cognitive processes as methods to study actual behaviour at an atomic level, the use of verbal reports and protocols for their analysis in data collection has evolved largely from efforts at understanding human problem solving in the field of psychology (Ericsson & Simon, 1980; Newell & Simon, 1972). Indeed, verbal protocol methods have been heralded as one of the most important methodological developments in psychology (Crutcher, 1994). Such methods have also been useful to studies of decision making processes and consumer judgment (Todd &

Benbasat, 1987, 1993). However, we have found that their particular value in information systems research is understated. This paper addresses that by describing an implementation of two forms of TAP implemented consecutively in an IS scenario. This section discusses the array of related verbal protocol methods described in consumer behaviour and psychology literatures; and explores their limited use in IS research. The sections following describe the specific protocols implemented in this study and display their value to IS research.

Many of the data collection methods involving the use of verbal reports are described quite similarly, though they can display subtle differences. An early example of the think-aloud method by Lewis (1982) at IBM research, sought to illuminate how people were interacting with computer interfaces. In think-aloud protocol (TAP) sessions, the participant is asked to vocalise their thoughts, their actions, their expectations of the results of those actions and, any confusion or concerns arising, in the presence of a facilitator. This facilitator also observes the session and may prompt the participant in order to keep the commentary alive. Dillman et al. (2000; 2008) describe a thinking aloud cognitive interview as the prevalent method for pretesting questionnaires. In this method the interviewer requests a respondent to proceed through a draft questionnaire while speaking the questions, and their cognitive processes in forming their responses, aloud, in the presence of the interviewer (Dillman et al., 2008, p221). The interviewer may intervene and probe the respondent to discover whether the intention of the questions is being accomplished. Lewis (1982) makes careful distinction between an observer and an interviewer. During a TAP session, an observer may *prompt* the participant to discover what the participant is thinking, however an interviewer may *probe* a participant to discover why they have attempted an action or what they believe an action will result in. In order to minimise observer influence on participants' cognitive processes during the TAP, Lewis cautions that probing interview questions should be saved for a post-TAP debriefing. A list of such *prompting* questions for use during the TAP session and *probing* questions for use later debriefing are listed in Table 2-1 below. The relation of the *concurrent* protocol to the TAP session itself and the *retrospective* method to a later debriefing session are discussed in the section following.

Prompting questions (Concurrent Protocol)
What are you doing now?
What are you looking for now?
What are you thinking here?
What do you want to do here?
What do you think should happen if you click that?
What has the system done now?
What is the system trying to tell you with that message?
Why has the system done that?
What were you expecting to happen then?
Why do you think that's confusing?

Probing questions (Retrospective protocol)
Did any of the questions now make sense? Which ones, and why?
Did any of the questions not apply to you? Which ones, and why?
Did the task instructions make sense?
What would you change in order to make the task easier?
Was the task real enough?

Table 2-1 Prompting a probing questions used during the TAP sessions (Developed with reference to Monk, Wright, Haber, & Davenport, 1993)

Within the family of verbal protocols, a number of related methods exist. The concurrent think-aloud protocol proceeds by requesting that participants vocalise their thoughts and actions while doing the tasks under study. This may be contrasted with the retrospective think-aloud protocol (also called *post-hoc* or free recall protocols (Kuusela & Paul, 2000)) wherein the thoughts and actions are verbalised after the task completion session has ended. The former has the advantages of a real-time commentary on the participant's intentions and reasoning during the event but is disadvantaged by the same measure in that such concurrent verbalisation of action and intent may influence those actions (Biehal & Chakravarti, 1989; Kuusela & Paul, 2000). The latter is often stimulated by a visual recording of the event itself such as screen capture recording for computer based tasks. The retrospective approach has the advantage that it can be facilitated after a more normal progression of actions during the event, however it suffers from its retrospective nature allowing participants more time to reason out their actions, constructing logic that may not have actually existed (Biehal & Chakravarti, 1989). Prior studies have found that while the concurrent method outperforms the retrospective in terms of the amount of data collected, the retrospective protocol offers more information on the later stages of the subject task (Kuusela & Paul, 2000). As such the concurrent protocol may be distinguished from the retrospective protocol in terms of the stage at which it occurs as implied by the protocols naming. In addition to that, the protocols may also be distinguished in terms of what the participant is asked to do and what the session facilitator attends to during each stage. These distinctions are diagrammed in Figure 2-1 below.

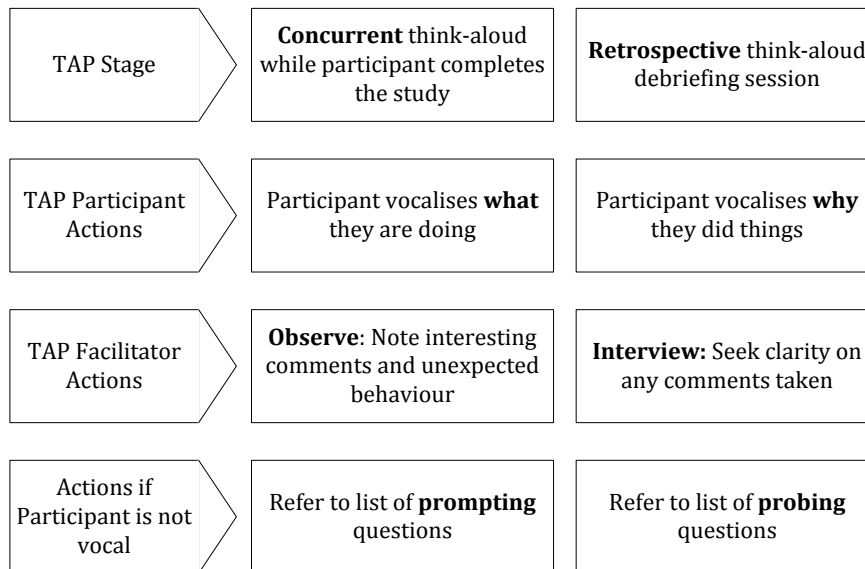


Figure 2-1 Comparison of concurrent and retrospective TAP methods

Our study used both methods at different stages during the TAP sessions. While each participant worked through a shopping decision task, the facilitator operated the session under the principles of concurrent TAP. Afterwards, a debriefing session was run using the principles associated with retrospective TAP. As such, this paper describes the dual use of the protocols during a single study. In addition, the protocols were employed to serve two distinct goals, namely (i) to test the propriety of the task and software for the study, and (ii) to test the face validity of the constructs in the questionnaire.

3 Think-Aloud Protocol

Many IS studies test information systems through user tasks and collect data through user questionnaires. Such research scenarios are particularly suited to TAP as a pre-testing of both the instrumentation and the tasks users complete in information systems. In studies where such data collection involves an online questionnaire or new software, pre-testing with TAP studies has the added advantage of testing the questionnaire and task software. Here we describe the use of TAP studies to verify questionnaire data collection instruments and user tasks in terms of participant apprehension as well as technical achievability.

3.1 Background to the TAP sessions

The implementation of TAP sessions formed a central part of a larger research project, diagrammed in Figure 1-1 earlier. The project consisted of two earlier pilot experiments where participants proceeded through a shopping decision task using software developed for the research supporting three treatment groups in the shopping scenarios. After these pilot studies, that software was further developed and more shopping scenarios were introduced. The research plan involved a call for participation in a main experiment through a partner online shopping company. In order to increase the technical viability of the software and the validity of the online shopping task,

instructions, and data collection instrument, a series of twelve double TAP sessions were conducted before the software was released for the main experiment.

Data were collected through an online questionnaire after participants had navigated through a shopping decision task in an online shop created for the purposes of the study. Prior to the shopping decision task, participants were provided with a pre-task questionnaire and instructions as to the task through the same software. Further data were collected in the background, such as the participant's screen-size, browser signature and session duration. Before launching the software in a main field experiment, the efficacy of the software for task and questionnaire completion as well as participant apprehension of the instructions, task and questionnaire items were tested using the double TAP method.

The pre-task questionnaire gathered demographic information such as age range, gender and previous online experience. The second questionnaire gathered data on the variables of interest in the study and was administered after participants had completed the tasks. For all TAP sessions the software was accessed from the server on which it would run for the main experiment. The shopping decision task, questionnaires, and the instructions for the study were all facilitated by software created specifically for the study. Researchers have encouraged the use of questionnaires combined with other research methods in order to develop richer understanding of results (Pinsonneault & Kraemer, 1993).

Twelve think-aloud sessions were conducted on the software which was used for the later main experiment. Traditionally eight sessions are conducted per test, with as few as three or four recommended in more recent implementations (Krug, 2009). In consideration of this and the data attained, we decided after twelve sessions that little could be gained from conducting any further sessions.

The software was installed on a web server used by the partner company. For the main experiment the software was accessed from this same web server. Running the software for the think-aloud sessions from the location which was to be used for the main field experiment allowed testing *in-situ* such that any errors of connection or server compatibility issues could also be investigated through the TAP sessions.

Access to the web server was provided for each think-aloud participant using the same laptop, however each participant was provided with the operating system and browser which they would normally use. This was achieved by having a series of virtual machines and multiple browser installs on the same laptop. Participants were also asked whether they preferred using a mouse or the track pad, and their preferred choice was provided to them. All except one of the think-aloud sessions were conducted in a quiet teaching room within the researchers' university. The other was conducted in the participant's own office.

We prepared a number of documents to ensure the smooth running of each think-aloud session. A briefing document explained the method, the task and sought the formal consent of each participant. We took notes on comment sheets created for each session. As recommended by Monk et al. (1993), we created a list of questions to prompt continued vocalisation should a participant remain quiet for a time. A debriefing sheet was also composed to address questions which were expected to be useful should they have not been answered in the normal course of the session. This approach is in line with Lewis' (Lewis, 1982) distinction between *prompting* the participant in order to

continue their monologue during the TAP session while leaving *probing* questions for a debriefing session. The first part of the sessions conducted used only prompting questions, consistent with descriptions of *concurrent* think-aloud protocol. Later, in a debriefing stage, participants were asked more probing questions consistent with *retrospective* think-aloud protocol (Biehal & Chakravarti, 1989; Kuusela & Paul, 2000). The set of documents and the associated protocol for running the sessions were created with reference to Monk et al. (1993), Krug (2000; 2009) and on advice from a number of senior colleagues who had previously used versions of TAP methods.

3.2 Running the TAP sessions

At the beginning of each session, each participant received a copy of a briefing sheet explaining the purpose of the session. Participants were also asked for permission to record the audio from the session. This sheet requested a participant signature to indicate their consent. During each session, we took note of any interesting comments and any unexpected behaviour observed. A list of prompts was referred to by the researchers whenever the participant became silent, in order to restart their vocalisation of thoughts and actions, e.g. ‘What are you doing now?’

As each participant proceeded through the initial instruction screens they were asked to vocalise the text as they read. Any area of text which they read twice or stumbled on was noted by the researchers. In addition to vocalising the instructional text displayed on screen, participants were asked to vocalise each action they took, and narrate their thoughts aloud. The briefing document asked that participants be particularly talkative on aspects of the software that were not clear or straightforward or that caused them to make mistakes or become confused. Each participant was informed specifically that the software was under scrutiny and that they were not. This protocol was followed by participants through a pre-task questionnaire, the on-screen instructions, the task screens and a final questionnaire, presented after the tasks had been completed. In particular at the final questionnaire stage, participants were encouraged to read each questionnaire item aloud as they encountered it. They were encouraged to vocalise their answers and their thought processes while formulating their response. If a participant indicated difficulties, stumbled on an item, or showed signs of confusion, we made a note to review the item.

After reviewing the systems, a short debriefing session was conducted to discuss any notes the researchers had made during the review. This formed the retrospective stage of the double TAP sessions and was guided by a debriefing document. During the retrospective stage, we enquired directly of each participant whether the instructions, tasks and questions had made sense to them and whether there was any grammar, wording or language which they found strange. The questions asked in this part of the session were more probing than the prompting questions asked during the concurrent stage, e.g. ‘Did any of the questions not apply to you? Which ones, and why?’. Any questions arising from researcher notes taken during the concurrent stage were addressed. This distinction between the kind of questions asked at each stage is in line with Lewis’ (Lewis, 1982) suggestions on the conduct of TAP sessions. The debriefing stage was also used to enquire further from participants as to any issues recorded during the earlier stages. In this way, anything which produced difficulties or confusion for the participant was investigated and recorded for later review. The audio recordings taken during each session were reviewed within a short time of each session. The notes, taken

during each session were also reviewed. These were then collated into a spreadsheet listing any issues which arose during the sessions to be addressed. The number of times each issue arose in the TAP sessions and any specific comments made by participants were also noted. Dillman et al. (2008, p225) warn of the danger in making changes, in particular with regard to questionnaires, based on the experience of one, or few respondents in TAP sessions. Those concerns were addressed in the procedures used for this research by discussing each issue with a senior colleague. Each item was discussed in turn to determine whether alterations should or should not be made, while remaining cognisant of the possible over-influence of individual cases.

4 Discussion: implications of the TAP sessions

The most notable change made to the software due to the TAP session data was a reduction in the amount of instructions presented to participants. In the earlier screens, instructions had been displayed describing how to use the software and how to proceed through the task. A number of participants were observed to skip the instructions screen, skim over the text or vocalise that they would not normally pay much attention to such a screen if they were not participating specifically to help find issues with the software. One participant in particular remarked: *"I'm reading it pretty carefully but I'd never normally read anything this carefully online unless it was... you know, something to do with my job but when I'm shopping around I just click around and instantly expect to know, and feel familiar with what I have to do"*. After deliberating this, we decided to reduce the length of the instructions. These screens were reduced to only include an instruction as to the purchase decision itself, and basic instructions as to how to complete the task with the software.

It became apparent in the debriefing sessions that some of the participants did not distinguish easily between the screens which were instructional and which were specifically shopping task screens. To distinguish these types of screen better, the background colour of the shopping task screens was changed. This change also allowed participants to more easily remember the shopping task screens when answering the questionnaire.

A number of constructs of interest to the study were measured using questionnaire items. This paper focusses on the effect of the TAP sessions on the evolution of the questionnaire. The questionnaire was presented in the software after the shopping task screens. Each of the constructs in the larger study was measured via a series of questionnaire items. A number of participants expressed frustration at being asked ostensibly the same question more than once in the questionnaire. The TAP sessions allowed the researchers to identify the items which participants considered to be redundant and remove or rephrase them as appropriate. This had the added advantages of reducing the size of the questionnaire and its overall completion time for participants. Three of the constructs measured in the larger study were reduced from six-item constructs to five-item constructs. The same constructs had been tested prior to the TAP studies in pilot experiments. The Cronbach alpha scores for the constructs had been acceptable at the pilot stage, and remained acceptable when tested in the main experiment, after the TAP studies with the reduced set of items. This enabled our larger study to proceed with a shorter questionnaire without losing reliability. This displays an added advantage of using TAP sessions before a large scale questionnaire is launched.

A series of smaller changes also resulted from the TAP data. The TAP studies allowed for the expected duration reported on the call for participants for the main experiment to be based on actual data. Additionally, though the main purposes of the TAP were pre-testing the software itself and ensuring face validity of the questionnaires, the TAP provided an early insight into what was happening. One participant remarked that he would always seek product reviews, the content of which would always outweigh price differences and quality signifiers on a vendor website. A number of the TAP participants were observed to scan the product data set but still compare on some specific attributes at the same time. One participant remarked "...comparing them now by [one attribute] first... scanning the [another attribute] as well'. Through operating the TAP sessions, the researchers gained some early insight into the behaviours which participants exhibited.

5 Conclusion

This study demonstrates the double-value of think-aloud protocols in terms of pretesting the software systems and associated instructions used in the main experiment, and its use in pre-testing the questionnaire. As such, the study displays the benefit of using a mixed method approach in this context.

The study also displays one way in which information systems researchers may combine quantitative methods with more qualitative approaches such as think-aloud protocols. The combination of laboratory experiments, think-aloud protocols and main field experiment allowed for a more comprehensive approach to the research questions than either method alone. The value of the TAP sessions was explicit, in facilitating review of the instruments used, providing clarity to users on the task, and indeed providing clarity to the researchers on operational aspects of the study. The particular implementation of the TAP sessions also provided a distinct contribution to effecting a robust and rigorous overall study: concurrent TAP provided most value in assessing the propriety of the task and software for the study, and retrospective TAP provided value in assessing face validity of the questionnaire constructs.

Overall, this paper illustrates not only the value of a combined implementation of two distinct types of TAP in the context of an online consumer behaviour study, it also contributes to behavioural information systems studies by contextualising how TAP can provide a bridge between two variants of experimental methods. In particular, this paper demonstrates how a quantitative pilot-based laboratory experimentation can be improved using a more qualitative intermediary step based on TAP, to directly facilitate a further quantitative field experiment.

Acknowledgement

This work was supported, in part, by Science Foundation Ireland grant 10/CE/I1855 to Lero - the Irish Software Engineering Research Centre (www.lero.ie), and in part by PRTL I IV.

References

Biehal, G., & Chakravarti, D. (1989). The Effects of Concurrent Verbalization on Choice Processing. *Journal of Marketing Research*, 26(1), 84-96.

- Crutcher, R. J. (1994). Telling What We Know - The Use of Verbal Report Methodologies in Psychological-Research. *Psychological Science*, 5(5), 241-&.
- Dillman, D. A. (2000). *Mail and Internet Surveys, The Tailored Design Method* (Second ed.). New York: John Wiley and Sons, Inc.
- Dillman, D. A., Smyth, J. D., & Melani, C. L. (2008). *Internet, Mail, and Mixed-mode Surveys: The Tailored Design Method* (3 ed.). New York/US: John Wiley and Sons Ltd.
- Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review* 87(3), 215-251.
- Krug, S. (2000). *Don't Make Me Think! A Common Sense Approach to Web Usability*. Berkeley, CA: New Riders.
- Krug, S. (2009). *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems*. Berkeley, CA: New Riders.
- Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology*, 113(3), 387-404.
- Lewis, C. (1982). *Using the "thinking-aloud" method in cognitive interface design*: IBM T.J. Watson Research Center.
- Monk, A. F., Wright, P. C., Haber, J., & Davenport, L. (1993). *Improving Your Human Computer Interface*: Prentice Hall.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Pinsonneault, A., & Kraemer, K. L. (1993). Survey research methodology in management information systems: an assessment. *Journal of Management Information Systems*, 10(2), 75-105.
- Todd, P., & Benbasat, I. (1987). Process Tracing Methods in Decision Support Systems Research - Exploring the Black-Box. *MIS Quarterly*, 11(4), 493-512.
- Todd, P., & Benbasat, I. (1993). An Experimental Investigation of the Relationship between Decision-Makers, Decision Aids and Decision-Making Effort. *Infor*, 31(2), 80-100.