Getting the Whole Picture: Collecting Usability Data Using Two Methods—Concurrent Think Aloud and Retrospective Probing

Julie H. Birns, Kristen A. Joffre, Jonathan F. Leclerc & Christine Andrews Paulsen
American Institutes for Research
kjoffre@air.org, cpaulsen@air.org

Abstract

Two data collection methodologies used in usability studies are 1) concurrent think aloud and 2) retrospective probing. This paper draws upon research and practical experience to define the strengths and weaknesses of these methodologies, including the role of memory in each process, and the value of using both methods together.

Introduction

The cornerstone of usability testing is user feedback. Various methods are utilized to derive information from users about their experiences with an interface. One effective way to obtain information is to simply observe users' behaviors while they work with an application. When usability issues exist within the interface, users often hesitate, struggle, or become frustrated. Observing user behavior supplies critical information about a wide range of variables, such as workflow, navigation and terminology.

However, observing behavior does not reveal exactly what users are thinking or perceiving. This information is attained only through eliciting users' verbal feedback. We frequently use two methods for obtaining user feedback: concurrent think aloud and retrospective probing.

Concurrent Think Aloud Procedure

What is it?

Concurrent think aloud is used extensively in cognitive psychology to learn how individuals process information while performing complex tasks. In relation to usability testing, concurrent think aloud refers to users verbalizing thoughts while interacting with a system (Dumas & Redish, 1999; Ericsson & Simon, 1993; Halpern, 1989; Nielsen, 1993; Wickens & Hollands, 2000). During this process, users discuss their actions, perceptions, and expectations regarding the application's interface and functionality. We have learned a lot from think aloud exercises, including information about users':

- level of comprehension of a system's purpose and functionality,
- initial expectations of where features are located within a system's interface,
- reactions to the visual design of an interface, and
- color preferences.

What are its strengths?

Concurrent think aloud is an effective method for collecting information regarding users' experiences. It enables administrators to identify where users are in a series of tasks, follow their thought processes, and identify points in the task flow where users deviate from the ideal path.

As users are providing feedback while completing a task, this method vividly reveals users' conceptions and misconceptions regarding a system. In contrast with other methods where the test administrator must interpret users' actions, with the concurrent think aloud method, usability issues are disclosed as they naturally occur in the workflow. For example, if users were given the task of ordering an item via an on-line consumer shopping website, users applying the concurrent think aloud protocol could verbalize that they expect to click on the "shopping cart" icon to checkout, rather than the "checkout" button.

The feedback obtained using concurrent think aloud is "real-time". In other words, because of the concurrent nature of the procedure, users do not have time to rationalize their thoughts (Ericsson et al., 1993; Nielson, 1993). This presents a fairly unobstructed view of the users' actual experiences. Moreover, these reactions remain uninfluenced by subsequent experiences with the interface. For example, we are able to capture information before users either blame themselves for misinterpreting a system's terminology or reassess its meaning. The think aloud protocol reveals their initial, candid reactions and/or understanding of the terminology's meaning.

Further, the nature of the feedback allows test administrators to immediately identify specific aspects of the interface that the users perceive as positive or negative, easy or difficult. As a result, the test administrator understands what users are thinking at that moment and is able to immediately adapt to an appropriate line of questioning that addresses the issues users verbalized.

What are its weaknesses?

Despite the candid nature of the feedback acquired during thinking aloud, the procedure is not completely natural because the act of verbalizing thoughts while completing a task can potentially impact the normal workflow (Ericsson et al., 1993; Nielsen, 1993). Typical users of a computer application do not usually verbalize their thoughts while completing tasks, such as opening a file or locating the print function.

Verbalizing perceptions and observations causes users to process information on multiple levels (e.g., visual and auditory) (Seamon, 1980). In particular, processing information at both the visual and verbal levels may impact users' problem solving strategies. While describing their behavior and perceptions, users may be more likely to notice inconsistencies in their mental models of the system. As stated by Nielsen (1993), "The dual processing may make them more aware of illogical or incongruent observations and statements." Furthermore, verbalizing thoughts while performing tasks requires a great deal of attention and concentration. Verbalizing may cause users to attend to the application and task components more precisely than they would if they were not doing so. In their natural environment, users may not direct their full attention to the task and application, thus decreasing the level of mental energy and problem solving devoted to understanding the interface.

In addition, thinking aloud concurrently can impact performance and productivity data, as measured by task time (Ericsson et al.,1983). A study conducted by Berry & Broadbent (1984) compared user performance, as measured by task time, on tasks where users were thinking aloud versus working silently. This study revealed that users performed tasks 9% faster if they were asked to use the concurrent think aloud procedure. As such, one cannot gather accurate performance data when users are thinking aloud concurrently while completing tasks.

Possible Memory Constraints

During the concurrent think aloud method, users attend to multiple novel stimuli. These include the application's features, an unfamiliar environment, and the test administrator and observers (if present). As various attention models suggest, humans are not capable of simultaneously attending to all stimuli present in their environment (Glass, Holyoak & Santa, 1979; Halpren, 1989; Seamon, 1980). If a feature on the application is not attended to, users will not be able to recall information regarding that feature. Further, they have the added task of verbalizing their thoughts. This increases the users' cognitive load, and, consequently, may impede their ability to attend to the interface's features.

Conclusion

Concurrent think aloud reveals an abundance of information about users' experiences, thoughts, and perceptions regarding an interface. Despite its limitations, concurrent think aloud is the best methodology the authors have found for collecting real-time cognitive data during usability tests.

Retrospective Probing Procedure

What is it?

Retrospective probing involves asking users a series of questions about their experience with a system immediately after they have completed a task or series of tasks with a system (Dumas et al., 1999; Ericsson et al., 1993; Nielsen, 1993; Wickens et al., 2000). These questions reveal the users' memories of their experiences with a task or system. Often, users' responses will

highlight major usability concerns or issues that are prominent in the users' minds. Retrospective probing questions can generally be presented in two forms, open-ended and close-ended.

Open-ended questions encourage the users to freely express their perceptions of the system. Although these questions are somewhat structured, a properly worded question will not influence users' responses. Open-ended questions result in qualitative data that describes user experiences. Examples of open-ended questions include the following:

- What do you like best about this application?
- What do you like least about this application?
- What aspects of the application, if any, would you like the designers to change?

Close-ended questions can range from those which elicit a simple Yes/No response to those which are comprised of a five- or seven-point Likert scale (see example below).

Overall, this task was easy to complete:

1	2	3	4	5
Strongly Agree				Strongly Disagree

Likert-type scales ask users to rate, or give a numeric value to, their experience, thus providing quantitative data. However, assigning a numerical rating to one's experience is extremely subjective. As such, we use close-ended questions as a base for further discussion. Understanding the reasons why users choose particular ratings often reveals more about a system's usability than the numerical rating. For example, users may rate the example above, this task was easy to complete, as a "5-strongly disagree." The test administrator would then prompt users to explain why they assigned this rating. Subsequently, the administrator may encourage a discussion of the interface's particular qualities that users found made the task difficult or confusing.

What are its strengths?

One of the fundamental strengths of the retrospective probing method is that it provides a holistic view of users' experiences. Essentially, it encourages users to express their responses and reactions within the framework of their overall experience with the interface. Likewise, the method encourages users to evaluate their experiences with respect to all tasks, not simply individual ones.

The questions asked during retrospective probing are typically consistent from task-to-task and user-to-user. This provides responses that are easily classifiable and comparable across both tasks and users.

Retrospective probing is an effective technique for acquiring user feedback because it renders thorough, reflective responses that highlight key points of users' experiences. This line of questioning allows users to assess what they perceived to be the most important strengths and weaknesses of the system's interface; it is a measure of users' *perceptions* of their experiences, which are often as revealing as the experience itself.

What are its weaknesses?

When reflecting upon their experiences, users often develop theories and rationalizations for their behavior, conceptions, and misconceptions that may not be accurate descriptions of their experiences. They may focus solely on these theories and rationalizations, rather than share their actual experiences and thoughts about working with the system.

Possible Memory Constraints

Retrospective probing requires users to recall past experiences. Limitations that occur during different stages of the memory process may impact users' abilities to accurately describe their experiences with the product (retrospectively). User tasks during an evaluation sometimes depict these memory limitations.

While attending to the multiple steps of a task, users may forget qualities of their experiences attempting previous steps due to their limited memory spans. Normal adults can typically recall, at most, seven, plus or minus two, digits, letters, or unrelated words (Ericsson et al., 1993). Each new task step interferes with the previous steps. By the time the users are questioned about their experiences, the first and most recent steps appear to them to be the most salient. Users often recall the first step due to the primacy effect, which is the tendency to recall items presented first in a series. Likewise, they often remember the last item due to the recency effect— the tendency to remember the most recent stimuli or task—as nothing else has interfered with it (Ericsson et al., 1993; Glass et al, 1979; Halpern, 1989; Seamon, 1980).

Limitations in memory retention may also be explained by the decay hypothesis, which asserts that information in the user's immediate memory will be forgotten, unless it is rehearsed (Ericsson et al., 1993; Glass et al, 1979; Halpern, 1989; Seamon, 1980). While completing new stages of a task, users do not have the attention or cognitive capacity to rehearse events from prior task processes. Consequently, users will likely forget important aspects of their total experience.

Errors in memory also occur during the retrieval process. Studies suggest that if an individual is under a high level of stress, such as during a test, there is a decrement in performance and memory. A usability test creates a unique stress that is derived from performance anxiety and the knowledge that others are observing one's behavior. If users experience too much stress, it may negatively impact their performance and memory (Glass et al, 1979; Halpern, 1989; Seamon, 1980). Thus, users may not recall specific information about their experience.

Conclusion

Retrospective probing is particularly adept at revealing users' overall perceptions of a task or interface. Despite its limitations, retrospective probing is the best methodology the authors have found for collecting usability test data that can be easily classified and compared across tasks and users.

Using both procedures

Why is it important to use both procedures when collecting user feedback data?

Both the concurrent think aloud and retrospective probing procedures have inherent benefits as well as limitations. In addition, each method provides unique feedback about users' experiences, one simultaneous with the users' actions and one upon completing a task or series of tasks. Users' experiences with a system are not simply the process of completing tasks, but also include their perceptions and memories of the experience. Thus, the authors have found that a combination of the two methods is the most comprehensive and effective way to attain insight into users' experiences and understand a system's usability issues. The benefit of utilizing both methods is that together they provide a more realistic view of the user's entire experience: thinking aloud offers insight into the user's thought process during the task, and retrospective probing reveals information about the user's recollection and conceptualization of the experience.

Conclusion

Concurrent think aloud and retrospective probing methodologies offer unique insight into users' true experiences, as, of course, the technology does not yet exist for researchers to "read" people's minds. Although limitations in memory and differential cognitive load may impact users' abilities to offer accurate and descriptive feedback during concurrent think aloud and retrospective probing, the information users provide during these processes is an extremely beneficial and necessary component of understanding usability issues. While it is impossible to know the users' entire thought processes and perceptions of their experiences, combining both methods enhances the assessment of the users' interactions.

References

Berry, D.C. & Broadbent, D.E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36A*, 209-231.

Dumas, J., Redish, J. (1999). *A practical guide to usability testing*. Portland, Oregon: Intellect Books.

Ericsson, K.A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, Massachusetts: The MIT Press.

Glass, A., Holyoak, K., Santa, J. (1979). *Cognition*. Reading, MA, Addison-Wesley.

Halpern, Diane F. (1989). *Thought and Knowledge: An Introduction to Critical Thinking*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Nielson, J. (1993). *Usability Engineering*. Chestnut Hill, MA: AP Professional.

Seamon, John G. (1980). *Memory & Cognition*. New York: Oxford University Press.

Wickens, C. & Hollands, J. (2000). *Engineering Psychology and Human Performance*. Upper Saddle River, New Jersey: Prentice Hall.