The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking

Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, Judith Ramey
Department of Technical Communication, University of Washington, Seattle
14 Loew Hall, Box 352195, Seattle, WA 98195-2195
{zguan, sllee, ecuddihy, jramey }@u.washington.edu

ABSTRACT

Retrospective Think aloud (RTA) is a usability method that collects the verbalization of a user's performance after the performance is over. There has been little work done to investigate the validity and reliability of RTA. This paper reports on an experiment investigating these issues using the method called stimulated RTA. By comparing subjects' verbalizations with their eye movements, we found stimulated RTA to be valid and reliable: the method provides a valid account of what people attended to in completing tasks, it has a low risk of introducing fabrications, and its validity is unaffected by task complexity. More detailed analysis of RTA shows that it also provides additional information about user's inferences and strategies in completing tasks. The findings of this study provide valuable support for usability practitioners to use RTA and to trust the users' performance information collected by this method in a usability study.

Author Keywords

Retrospective think aloud, validity, reliability, verbalization, eye tracking, usability research.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Think aloud (TA) is a usability evaluation method used to gain insight into how people work with a product or interface. In the most commonly used approach, Concurrent Think aloud (CTA), users work on typical tasks while at the same time verbalizing what they are thinking and doing. Following the appearance of Ericsson & Simon's milestone work [8], this method became widely used in cognitive science and human-computer interaction (HCI). In HCI, CTA has been widely used to study various materials from webpages [6, 16, 31] to end-user products [1, 7], and in various settings from the laboratory [14, 15] to the field [23]. As Jakob Nielsen commented, "think aloud may be the single most valuable usability engineering method"[17].

However, certain questions have been raised about CTA. First, the act of speaking concurrently may have a negative effect on users' task performance. Second, the effort that users make to verbalize information while performing tasks might distract subjects' attention and concentration. Third, the effort to fully verbalize the steps in the work might change the ways that users attend to the task components [3, 19, 20, 24].

To avoid these possible negative effects, some usability researchers have proposed to use Retrospective Think Aloud (RTA), a method that asks users first to complete the tasks and only afterward to verbalize their process. This method is also called post-task testing [29], retrospective protocol [28], retrospective report [9], think after [3], etc.

RTA has been widely used, and people do believe that it provides valuable data; however, there has been little work done to confirm the validity and reliability of RTA. Most of the research to date on RTA has focused on comparing this method to other methods (e.g., CTA) in specific task domains [2, 3, 5, 11, 18, 29, 30]. These comparisons were based on user testing rather than experimental study, which undermines the validity and generalizability of the conclusions drawn [13]. No research has scientifically studied the validity of RTA based on its most fundamental claim—that in RTA people talk about what they really did in terms of their actual mental processes or performance. Thus the validity of RTA in usability research is still in need of serious investigation.

In this paper, we present an experimental study with three main goals: (1) to assess the validity of RTA (whether people's report of what they did truly follows their original task performance), (2) to evaluate the impact of task complexity on the validity of the RTA, and (3) to characterize what other information the RTA provides beyond the basic record of task performance.

First, we present our hypotheses and the details of our experiment. Next, we describe our data processing and findings. Then, we discuss our results and their implications

CHI 2006, April 22-27, 2006, Montréal, Québec, Canada.
Copyright 2006 ACM 1-59593-178-3/06/0004...$5.00.

for usability evaluation. Finally, we make some concluding remarks and discuss future work.

HYPOTHESES AND QUESTIONS

The focus of our study is the validity of RTA—whether subjects' verbal accounts accurately reflect what occurred during the task performance. The subjects' RTA is considered valid if it describes the same sequence of objects in the same order as the subject attended to in the original task performance. We also studied the reliability of RTA across two levels of task complexity. We worked with what appears to be the most commonly used form of RTA, "stimulated" RTA, in which the retrospection is prompted by visual reminders of the tasks [2, 5, 21].

We investigated two hypotheses:

Hypothesis 1: *People's recounting of what went on in their task performance in a stimulated RTA describes the same sequence of objects in the same order that the subject attended to in the original task performance.*

Hypothesis 2: *The validity defined in hypothesis 1 is not affected by the task complexity, which is defined in terms of visual information processing complexity.*

We also looked at two more exploratory questions:

Question 1: *Besides a record of the items attended to in the order they were considered, what other types of information does stimulated RTA provide and in what format?*

Question 2: *What is NOT in the stimulated RTA?—What features of the task performance are not reported?*

Decomposition of Verbal Report

To address the four concerns listed above, we decomposed the verbal reports into two aspects. Aspect one is the simple record of the objects that subjects report attending to during the task performance and the order in which they did so. This part of the verbal report can be empirically measured and compared with other independent validation data, e.g. eye fixations. Hypotheses 1 and 2 deal with this part of the verbal report. We evaluate this aspect of the verbal report along two dimensions:

1) Degree of valid account: to what degree does subjects' retrospective verbalization truly report what they attended to, in order, in the task performance?

2) Degree of fabrication (error of commission): to what degree is the retrospective verbalization based on subjects' fabrication of events that in fact did not occur?

The measures of valid account and fabrication indicate the validity of stimulated RTA, as stated in hypothesis 1. Whether these two measures are affected by task complexity indicates the reliability of RTA, as stated in hypothesis 2.

The second aspect of the retrospection is *how* subjects talked about the objects that they attended to. Question 1 addresses this aspect of the verbal report. RTA can be most informatively studied by categorizing (1) the types of verbalizations that occur and (2) the way they are related to steps in the task performance sequence.

In addition to what is in people's retrospective verbalization, it is equally important to see what's *not* there, which in studies about TA [8, 24] is sometimes called forgetting or the error of omission. But not including certain information in the verbal report doesn't necessarily mean that people forgot what they did. They may simply choose to report information in a different way or in less detail. Hence, we can only interpret instances in which objects were missing in subjects' verbal reports as instances of *omission*, analysis of which answers question 2.

The results about the validity and reliability of stimulated RTA can be generally applied to any field that uses RTA to collect user's performance information. The results about the types of verbalization and missing information in stimulated RTA are more useful in the specific context of usability evaluation.

EXPERIMENT

We designed and conducted an experiment to capture and compare two records of the events that occurred during subjects' task performance: eye movement data and retrospective reports.

Use of Eye Movement Data as Validation Data

Eye movement data has been considered one of the measures or indicators of user attention [10] and has been compared with a record of people's concurrent think aloud [22]. It directly shows the locations that people have looked at and in what order. In our study, we used eye movement data as criterion data to indicate what objects people attended to and in what order. The logic of using eye movement data as criterion data is based on a generally accepted assumption called "eye-mind hypothesis" [12, 32] that where people look indicates what they are paying attention to, or thinking about.

Task Design

We designed our tasks as typical problem-solving tasks similar to the types of tasks that other researchers have used in evaluating verbal protocols [8, 24]. We designed the tasks to be experimental tasks instead of "real world" tasks in order to eliminate unwanted confounds and complexities in subjects' task performance, which could lead to difficulties in processing and analyzing subjects' verbalization and eye tracking.

We also designed the tasks with two different levels of complexity to address the issue of RTA's reliability. We designed four tasks, two in a "simple" group and two in a "complex" group. In each group there was one graphical task and one numerical task. Subjects worked on all four tasks.The answer key was randomized to multiple-choice options (A, B, C, etc.) in order to prevent bias due to subjects' knowledge of the solutions from previously-tested subjects. The tasks were:

Simple tasks: 1) number pattern (numerical): evaluate the sequence to identify the last number in that sequence; 2) matching puzzle piece (graphical): choose the correct puzzle piece that matches the target piece in the picture.

Complex tasks: 1) classroom data table (numerical): evaluate the maximum capacity and number of students to determine which term period shows the greatest overload; 2) bottle or airplane graph (graphical): analyze graphs to determine which graph best represents the height and volume of water poured into a container or to evaluate whether statements about the airplanes are true or false.

Task complexity relates to the cognitive load required in completing a task. For our task design we borrowed classic concepts from Campbell [4] and Wood [33] to develop a combined definition of task complexity in problem-solving tasks: (1) information load: amount of information the subject has to retain; (2) information diversity: dimensions of information that need to be accounted for; (3) information transformation: amount of recoding of information for meaning; (4) number of dimensions in a solution, and (5) number of task steps.

Hence, matching a puzzle piece is a simple task because it requires less information load, diversity, transformation, and so on. The subject needs to remember the shape and/or color of the target piece and mentally rotate a puzzle piece to the same orientation as the target piece. The number pattern is also a simple task because it requires only a linear or constant mathematical calculation.

The complex tasks required greater cognitive processing in all five measurements. For example, the classroom data table required that subjects calculate the difference between room capacity and number of students to determine maximum overload across three classrooms. The airplane graph required that subjects use and retain information about two airplanes from three separate graphs. The bottle graph asked that subjects mentally envision how water flows into a container (flask, funnel, bucket, etc.) and translate that into a graphical representation.

Although we designed the tasks to be experimental tasks, the problem-solving strategies that they call upon are similar to those used for tasks in the real world and for tasks designed for usability testing: deriving answers from data presentations, identifying items based on shape, etc. Thus, the design of these tasks enables us not only to scientifically control the study, but also to ensure that the results could apply to usability testing using real world tasks.

Procedure

The experiment had four sections: a pre-questionnaire, a task performance session, an RTA session, and a post-questionnaire. The experiment took about 45-60 minutes.

The pre-questionnaire asked about subjects' background and experience in eye tracking and in using the think aloud method. After administering the pre-questionnaire, we tested subjects to determine whether their eyes could be accurately calibrated (if not, we ended the study).

If the eye calibration succeeded, subjects were asked to complete four tasks, two from the simple group and two from the complex group, with their eye movement captured. Subjects were also randomly assigned into one of two conditions (a Latin squared task order of simple-complex or complex-simple). The computer screen and subject's mouse interactions were recorded using a screen capturing software.

Following the task session, we briefly explained to subjects the basic concepts of think aloud (TA) and asked them to apply these concepts in a TA practice, in which they were asked to verbalize while taking staples out of a stapler.

After the training session, the video of screen captures was played on the computer. The video showed subjects the task screens they had seen in the task session, the cursor positions and movements, and any selections they made. The video did not show the captured eye movement. Subjects were asked to report what they did and what they thought when they were doing the tasks. The use of a videotape as a stimulus for the RTA is documented in the previous literature[2, 8, 9, 18, 21, 30]. Subjects' verbalizations were recorded by using video recorders.

After they completed the verbalization, subjects were asked to fill out a post-questionnaire about their perceptions of task complexity and their experience in doing RTA.

Subjects

Forty-three student volunteers were recruited from an undergraduate engineering class for this study. They received class credit for their participation.

Among these students, one student was dismissed because his eye movement couldn't be calibrated. Fifteen students' eye movement data needed substantial adjustments and were thus excluded from the analysis reported in this paper.

Another two students were eliminated because of difficulties with their verbalization. The exclusion criteria were (1) subjects rated their language ability as "speaking English is very difficult and I can only partially express what I really want to say", and (2) the evaluation of the verbal reports showed that their verbalizations were unintelligible. Another student's data was randomly excluded to achieve two groups of equal size.

In total, 24 subjects, two females and twenty-two males between 19 to 33 years old, were included in the data analysis reported in this paper. None of them had experience doing RTA, although one subject had once done a concurrent think aloud.

Apparatus

The experiment was conducted using a Dell computer running under Windows XP. The computer is equipped with an eye tracking system from Eye Response Technologies which includes an eye tracking camera, an ERICA system for eye calibration, and a GazeTracker for data collection. Subjects' task performance was recorded using Camtasia software. Their verbalizations were recorded onto Sony digital video tape using a video recording suite.

DATA PROCESSING

Coding of Sequences in Verbalization and Eye Movement

Eye movement data provides a highly detailed record of all the locations that a user has looked at. Reducing this data to a density level that can be compared to verbal report presents a challenge [22]. Our approach involved computationally reducing the eye movement data for each task to an ordered sequence of "Areas of Interest" (AOI), qualitatively coding the verbal data to ordered sequences of AOI, and then applying a sequence alignment algorithm to compare the AOIs in eye movement and verbal sequences.

"Areas Of Interest" as Indications of User's Attention

Coarse-level and fine-level rectangular AOIs were defined for each task screen, based on the "chunks" that might be looked at or talked about separately. Coarse-level AOIs were defined as major screen regions (e.g., instruction, task problem, answer choice, task submission button). When a coarse-level AOI includes meaningfully distinguishable objects, it was further decomposed into fine-level AOIs. For example, in the bottle graph task shown in Fig.1, the screen is decomposed into 5 coarse-level AOIs: an instruction area (A), additional textual labels (B), a problem area (C), a solutions area (D), and a task submission button area (E). The problem and solution areas (C & D) contain graphics that a person can meaningfully speak about or point to separately while describing the task. Thus, these areas were further decomposed into fine-level AOIs (f through m). Table 1 lists the number of coarse and fine level AOIs for each task.

Number of AOIs	Coarse level	Fine level
Puzzle	5	9
Number pattern	8	18
Classroom table	5	12
Bottle graph	5	8
Airplane graph	9	25

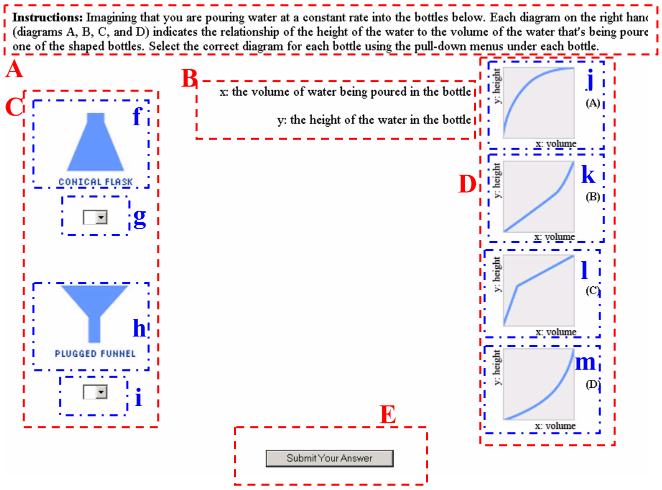
Table 1. Number of AOIs in the coding schema for each task

Coding of AOI Sequence from Eye Movement Data

Reducing the eye movement data to visual areas of interest involved the following steps: reducing the eye gaze stream to a sequence of eye fixations, determining which objects the users had fixated upon, and reducing the eye fixation data to AOI sequences. Once calibrated, our eye tracker is able to sample the (X, Y) screen location of an eye gaze 30 times per second.

Because we are specifically interested in the users' loci of attention, the eye gaze data first was transformed into a sequence of eye fixations (an eye movement that stabilizes an image directly on the retina for at least the minimum period of time required for processing the information). The GazeTracker software was used to calculate fixations, requiring a cluster of at least 3 gaze points within a 40-pixel diameter (slightly more than 1 degree of visual angle) for a minimum of 100ms for graphical and numerical data or 200ms for textual sentences (e.g., instructions). Assuming the eye-mind hypothesis [12, 32], the sequences of fixations represent the sequence of objects on the screen that the users cognitively attended to.

Because multiple fixations can occur in immediate sequence within one AOI (e.g., reading the instructions induces word-by-word fixations), any sequence of two or more fixations within the same AOI was collapsed into a single "fixation cluster." The generated sequences of eye fixations (clusters) were matched with the task screen AOIs to determine when fixations occurred within an AOI. This resulted in two eye movement sequences: a lower-resolution coarse-level sequence of AOIs and higher-resolution sequence that contained both coarse and fine-level AOIs (using fine-level whenever possible, but not all AOI regions decomposed at the fine level).



Instructions: Imagining that you are pouring water at a constant rate into the bottles below. Each diagram on the right hanc (diagrams A, B, C, and D) indicates the relationship of the height of the water to the volume of the water that's being poure one of the shaped bottles. Select the correct diagram for each bottle using the pull-down menus under each bottle.

x: the volume of water being poured in the bottle

y: the height of the water in the bottle

Figure 1: The coding schema for the bottle task. The coarse level AOIs are labeled in capital letter (A-E), and fine level AOIs are labeled with lower case letters (f-m)

Coding of AOI Sequence from Verbalization

The subjects' retrospective verbalizations were first transcribed into text files. During the qualitative coding process, coders identified utterance segments by categorical type and, when stated, the object AOIs that the segment referenced. The categorical coding of segments was based on a pre-defined set of verbalization categories which will be discussed later. When a subject verbally referenced an object in a task screen, such as "the conical flask" (see Fig. 1), the segment was coded at the fine level (AOI = "f") in addition to the coarse level (AOI = "C"). When a subject mentioned the region without indicating a specific object (e.g., "the bottle") or referred to the region as a whole (e.g., "the bottles on the left"), the segment was coded at only the coarse level (AOI = "C"). These coded segments were used to form the final AOI sequence for the verbalization.

Interrater reliability was calculated as the percentage of agreements out of the total number of codings per a verbalization session. It was performed on 16% of the data and yielded 87% reliability on the coarse level coding and 77% reliability on the fine level coding.

Calculating Validity Using Sequence Alignment

To measure whether subjects' verbalizations corresponded to the objects attended to in the order of occurrence, we compared the eye movement and verbal AOI sequences by calculating the edit distances and the alignment between two sequences using the Levenshtein algorithm and one of its extensions, the Needleman-Wunch algorithm. Levenshtein edit distance is a well-known algorithm for finding the minimum number of "edits" (i.e., deletions, insertions, or substitutions) required to transform one string into another [25]. The alignment of two sequences is a qualitative measure of the sequence similarity, which exhibits where the two sequences are similar and where they differ. In the HCI domain, Levenshtein distance has been used to measure error rates between the presented and transcribed texts in text entry[26], and to find out the missing or incorrect letters in cognitive modeling based on ACT-R model [27].

In this study, Levenshtein distance was used to compare eye movement and verbal AOI sequences on the coarse level. Given that the fine-level AOI verbal sequences could include both coarse or fine grain AOIs depending on the resolution that subjects used when referencing objects, Needleman-Wunch was used to allow for approximate matches. The sequence alignment algorithms calculated the number of "edits" to transform one sequence into the other, based on which maximal alignment of the verbal and eye movement AOI sequences was generated.

Once aligned, the AOIs from the verbal sequence that match up with AOIs in the corresponding eye movement sequence indicate valid accounts (the subject's verbal report corresponds to subject's performance.) The AOIs found in the verbal report but not in the eye movement data indicate verbal fabrication of information. Likewise, the number of AOIs found in the eye movement data but not in the verbal report indicates verbal omissions of information.

		Eye movement	
		Yes	No
Verbal report	Yes	Valid (approximate) account	Fabrication or misstatement
	No	Omission	N/A

Table 2. Measurement of "Valid Account," "Fabrication", and "Omission" based on the comparison of the verbal report and eye movement

We also found another feature of subjects' retrospection: misstatement. In this case, the subject mentions an object in-between two other objects reported in the eye movement data, but the subject misidentifies the middle object. Although the notion of misstatement doesn't appear in earlier literature, this case is different from fabrication and we make this distinction in our analysis. Table 2 summarizes the ways in which the alignments between the verbal report and eye movement data were compared. The results were normalized into percentages based on the total length of the verbal and eye movement sequences.

The following provides an example of alignment, which shows how we calculate the degree of valid account, fabrication, misstatement, and omission. Given the verbal AOI sequence: `BDBCGCF` and the eye movement AOI sequence: `ABCBACCEH`, the resulting alignment is:

```
Verbal report:    -BDB-CGCF-
                   |$| | |!
Eye movement:     ABCBAC-CEH
```

This sequence alignment shows that the verbal sequence consists of 4 valid accounts ("|"), 1 approximate account ("!"), 1 fabrication ("–" on eye movement sequence), and 1 misstatement ("$"). The total number of verbal AOIs is 7:

Degree of valid account = 4/7 = 57%
Degree of approximate account = 1/7 =14.3%
Degree of fabrication = 1/7 = 14.3%
Degree of misstatement = 1/7 = 14.3%

The total number of eye movement AOIs is 9. Five of them correspond with verbal AOIs. Omission accounts for the rest of them ("–" on verbal sequence.) Thus,

Degree of Omission = 4/9 = 44%

Categorization of Verbalization

As stated earlier, subjects' verbalizations about the objects that they attended to were coded to form the AOI sequence. In addition, subjects' verbalizations were coded based on what *kind* of statements they provided.

Earlier, Russo coded concurrent verbal statements into five categories: perceptual, low level inferences, high level inferences, strategy, and all others [24]. After a preliminary analysis indicated the presence of a broader range of categories, we coded our verbal reports into eight categories separated into four types, as shown in Table 3.

Type	Category
Behavior Statements	Procedural Behavior (PB)
	Negative Behavior (NB)
Inferential/explanatory Statements	Logic Inference (LI)
	Perception Explanation (PE)
	Strategy Explanation (SE)
Reflective comments	Forensics/Diagnostics (FD)
Others	Meta-Comment (MC)
	Others (OT)

Table 3. Categories of verbal statements

Behavior statements are specific statements about what subjects did during their task performance, such as "I read the instructions at the top" (A23). Negative behavior was coded for statements provided in a negative way, such as "I also don't think I read the name of the flask" (A23).

The inferential statements include "logic inference" directly inferred from or generated based on information that users attended to, such as "I see that the top and bottom of this highlighted piece protrudes out" (B01); "perception explanation", such as "the picture is pretty bright" (A01); and strategy explanation about how subjects completed the task, such as "and this one I just started doing the subtractions…addition" (B09).

Reflective comments contain self-diagnostics about what subjects did or should have done, such as "for this one I was actually a little confused about what they were asking at first" (A10). The last category is "meta-comments", such as "this one (number table task) kind of took me by surprise" (A18), and unidentifiable verbalizations. The coding of the categories of verbalization has 77% interrater reliability (percentage of agreements) on 16% of the data.

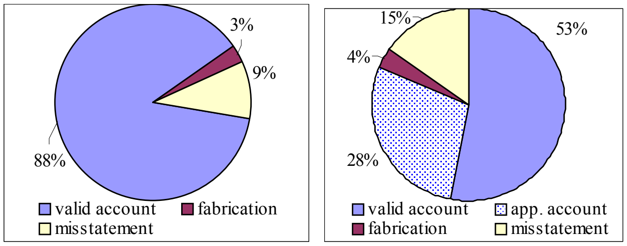
RESULTS

We analyzed the following data using descriptive statistics, repeated measure variance analysis, and chi-square:

1) Sequence comparison measures between verbal AOI sequences and ET AOI sequences
2) Percentage of categorical verbal statements
3) Subject's rating of task complexity and RTA experience

Validity of RTAP: Valid Account vs. Fabrication

Fig. 2 shows the validity of stimulated RTA on the coarse and fine level. Sequence comparison of verbal and eye movement data on the coarse level indicates the validity of RTA report on subjects' general problem-solving processes (Fig. 2-A). We found 88% valid accounts (verbal AOIs matched eye movement AOIs and occurred in the same order). 9% of misstatements points to subjects' awareness of having attended to AOI regions but inability to identify the exact target objects. And 3% fabrication in which verbal AOIs did not correspond with eye movement AOIs.



(A): coarse level **(B): fine level**

Figure 2: The validity of stimulated RTA

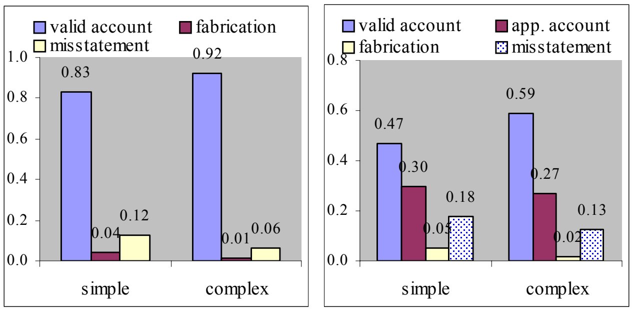
We also determined RTA validity at the fine level (Fig. 2-B). We found 53% valid accounts of low-level AOIs that matched up in the verbal and ET data; 28% of approximate matches in which verbal and eye data matched up sequentially on the coarse level but varied somewhat on the fine AOI levels (for example, a subject's verbalization may indicate the left side of a diagram, but the eye data the right side); 4% fabrication on the fine level; and 15% misstatement, indicating that people experienced some difficulties in identifying exact low-level AOIs even though they appear to remember attending to those regions.

We acknowledge that the design of the experimental tasks may lead to underestimates of fabrication because subjects are constrained to look at a defined set of items in a display. Any future study needs to look at the extent of subjects' fabrication in a real world task environment.

Reliability of Verbal Reports with Task Complexity

Our task design incorporated two levels of task complexity: simple and complex. To verify our measurements of task complexity, we relied on subjects' post-test ratings of task complexity on the four tasks that they worked on.

Subjects' ratings confirmed our measurements of task complexity. The repeated measure variance analysis shows significant difference between the two simple and two complex tasks (F(3,69)=13.948, p<.05.) A post-hoc Tukey analysis shows no significant difference between tasks in the simple group (the two puzzle tasks vs. the number pattern task, p=.572) and between tasks in the complex group (the classroom table vs. the bottle or the airplane task, p=.973). But there are significant differences between simple and complex tasks (p=.00, .00, .012, .003 for all four pair-wise comparisons). The results of subjects' rating show that tasks in the complex group are perceived as significantly more complex than tasks in the simple group.



(A): coarse level **(B): fine level**

Figure 3: The reliability of stimulated RTA over task complexity

To determine whether task complexity has any significant impact on the validity of RTA, we conducted a chi-square analysis of subjects' valid account, misstatement, and fabrication. We found that on the coarse level (Fig. 3-A) there is no significant effect of task complexity on the validity of RTA ($\chi_2 = 4.26, p = .12$). Subjects' valid account is 83% for simple tasks and 92% for complex tasks; fabrication dropped from 4% for simple tasks to 1% for complex tasks; and misstatement dropped from 12% for simple tasks to 6% for complex tasks.

Although there is no significant difference between simple and complex tasks on the validity of RTA, we did find an interesting trend: subjects tended to produce more valid accounts and commit fewer fabrications in the complex tasks than in the simple tasks. Could this suggest that subjects put more thought in complex problem-solving and can therefore verbalize in more detail? On the fine level (Fig. 3-B), we found no significant difference between simple and complex tasks (χ_3=3.6, p=.31) on valid account, approximate account, fabrication, and misstatement. The same trend that subjects produced more valid accounts and fewer fabrications on complex than simple tasks at the fine level is consistent with our findings at the coarse level.

Verbal Reports: Procedural, Inferential, and Explanatory

RTA also provides descriptive information about how subjects attended to the objects in their task performance.

We categorized the types of verbalization subjects provided and the way these verbalizations are related to the steps in the task performance sequence. We discuss the results on omission after presenting the results for verbal categorization, because we think there is a very close relationship between the two. Understanding the former could provide more insight into how omissions occur.

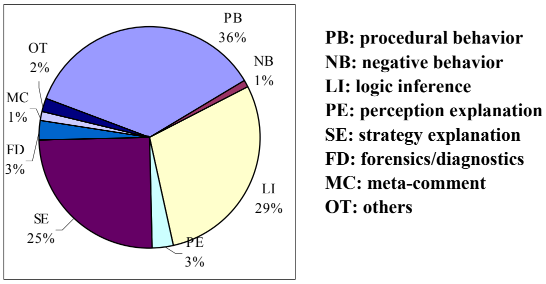


Figure 4: Categories of retrospective verbalization

Fig. 4 shows the distribution of the categorized verbalizations. 36% were statements about subjects' procedural behavior, 29% logical inference, 25% strategy explanation, 3% perception explanation, 3% forensics/ diagnostics, 1% meta-comments, and 2% for other. The low percentage of meta-comments and other statements indicates that subjects were focused on verbalizing what they recalled about their performance and that there was little intervention between the experimenter and subject.

A chi-square analysis of the verbal categories shows no significant effect of task complexity on the type of statements that people made (χ_7=3.69, p=.81). Fig. 5 shows that in both simple and complex tasks one-third of subjects' statements involved procedural behavior; 34% of logical inference for simple tasks and 24% for complex tasks.; 22% of strategy explanations for simple tasks and 28% for complex tasks; and 1% of other comments for simple tasks and 3% for complex tasks.

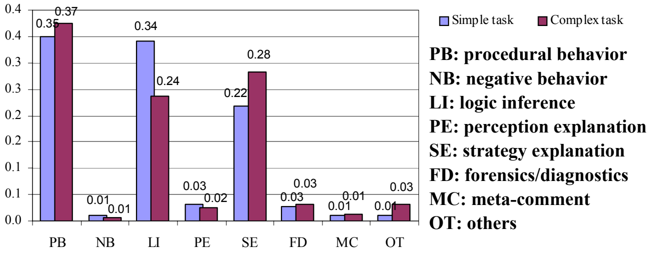


Figure 5: The percentages of categorized RTA statements

The results indicate that when subjects verbalized on complex tasks, they tended to make more higher-level inferences (strategy explanations) than intermediate-level inferences (logical inferences).

In terms of the relationship between the type of verbalization and reports of the objects that subjects were attending to, 41% of verbal AOIs came from procedural statements and 31% from statements of logical inference. Among the other verbal categories, 23% of verbal AOIs were drawn from explanatory statements, 2% from perception explanation, and 2% from forensics/diagnostics.

Degree of Omission

We were also interested in finding out what is *not* there, the degree of omission as revealed by the sequence alignment. We found that 47% of eye movement AOIs did not correspond with verbal AOIs. An analysis of the effect of task complexity on omission indicates significant difference between simple and complex tasks ($F(1, 23)=20.6$, $p<0.05$): 38% of eye movement AOIs were missing for simple tasks and 56% for complex tasks. There is no significant difference between tasks in the same task-complexity level.

Further analysis of typical omissions between AOI sequences in verbalization and eye movement suggests at least two possible reasons: One, differences in data density and abstraction level for verbal and eye data result in omission in general; and two, omissions more likely occur when subjects have difficulty working out a problem, which may explain why there are more omissions for complex tasks. We discuss omission further in the discussion section.

Subjects' Evaluation of RTA Experience

Subjects' rating of factors that facilitated their verbalizations (5: very helpful; 1: not helpful at all) shows that they relied on their memory the most (4.17), followed by the video replay (3.83). Video replay of their mouse movements (2.33) and the think-aloud training (2.13) were rated as not helpful. Rating of experimenter's prompts fell between being helpful and not helpful (2.61).

DISCUSSION

Hypothesis 1: *Our findings support our first hypothesis that people's recounting of what went on in their task performance in a stimulated retrospective think aloud describes the same sequence of objects in the same order as what they attended to during the original task performance.*

More than 80% of subjects' verbalizations of what they were attending to corresponded with the eye movement data. We reject the notion of subjects' fabrication since only less than 3% of their verbalization failed to match up with objects identified by their eye movement.

This finding indicates that usability researchers can trust the information they get from a stimulated RTA. This finding is especially useful for those whose products cannot easily be tested using concurrent think-aloud (for instance, games). Also, by using RTA researchers can collect other usability measures during task performance, such as time on task, error rate, etc., without concerns about the effects of verbalizing on that data. The combination of performance measures and verbalization can provide usability evaluators more accurate and comprehensive usability measures on the materials they tested. These gains are achieved at the cost of the additional time required for the retrospection.

Hypothesis 2: *Our findings support our second hypothesis that retrospective think aloud is reliable in that it is unaffected by task complexity.*

Subjects' verbalization on complex tasks, defined by a heavy and diverse information load, had the same percentage of valid accounts as their verbalization on simple tasks. In addition, the small incident rate of fabrication for complex tasks was similar to that of simple tasks. These results suggest the general reliability of stimulated RTA in usability testing, in which it is common to use tasks with different levels of complexity to investigate usability issues.

What other Information does RTA Provide?

Subjects' retrospective verbalization provided a wealth of explanatory information about what they were attending to, how they processed information, and how they arrived at a solution, and it did so while at the same time closely following the contours of the actual task performance.

Ericsson & Simon[8] considered explanatory statements as unreliable because they could distort the report of what subjects actually did and in what order. However, our study of what subjects attended to and in what order found that fully 23% of all verbal sequences used to correlate with AOIs in subjects' eye movement came from explanatory statements. Overall, only one-third of subjects' verbalizations were simply procedural and more than half were inferential (logical or strategic). While subjects' inferential and explanatory statements were not as specific as their procedural statements, they nonetheless provided important information about how subjects were mentally processing information to work out a solution.

What are People Omitting from RTA?

Our study found gaps in the verbal AOI sequences when compared to the eye movement AOI sequences, suggesting that subjects' were omitting information from their verbalization. To account for these omissions, we looked at what subjects were neglecting to say in their verbalization and arrived at two plausible explanations:

Case #1: Different data densities and levels of abstraction

Omissions occurred in part because verbal and eye movement data differ in data density and abstraction levels. Whereas eye tracking provides high density, low abstract-level sequence data, verbal reports tend to provide low density, high abstract level, aggregated sequence information. We anticipated this problem and tried to remedy it by using coarse and fine level AOI coding schemas. However, we found that the gap between RTA and ET could not be completely bridged in these instances.

To illustrate our point about data density and abstraction levels, we pick one representative case from our data, Subject B15 who had a total of 49 omissions (with .62 degree of omission), considered average across all subjects. In the verbal report on the second complex task (Fig. 1), which involves identifying the correct graph for the ink bottle, Subject B15 mainly talked about the ink bottle and the A, B, and D answer choices, and described his behavior: "…and I was pretty much looking from left to right the entire time; I glanced up at the instructions a few times…" We coded this part of the verbalization (`HHKDJMIA)`, following the coding schema shown in Fig. 1.

In contrast with the verbal AOI sequence, the eye movement AOI sequence was longer and richer in detail (`HKHKJMJHMKJBJAMJMHMHMLJKLIMIHM`). The codes H, J, K, and M appear multiple times in the sequence, which indicate that the subject's eye movement was constantly switching between the ink bottle and graphs A, B, and D.

Although in the alignment of the verbal and eye movement sequences 22 omissions were recorded, we could not simply dismiss them as a failure of the subject to report what s/he did. Subject B15 clearly stated looking left and right the entire time. Rather than repeating each instance of recursive behavior, the subject apparently chose to summarize his/her actions. Hence, this is one instance in which the eye tracking recorded the subject's recurrent eye movement between multiple information points but in which the verbal report reduced the ocular behavior to a single observation.

This difference in data densities and abstract levels could result from several facts, including prior training in RTA, auto-processing, etc. In the training session, subjects were told to verbalize everything that they were doing and thinking about. However, the subjects may be unsure of how much detail to provide. They tended to report on things that directly related to the task, such as selecting a choice, but were less likely to report the auto-processing steps, such as recognizing that the letter for the first choice is A.

Although different data densities and abstraction levels between verbalization and eye movement increase the number of omissions, we do not believe that this particular type of omission undermines the validity of RTA.

It is worth pointing out that when usability evaluators analyze RTA, they not only study what users are attending to but their behavior patterns. Given that the eye tracking gives credence to users' verbal report of their behavior, usability professionals can perhaps correlate specific behavior patterns with specific design problems (such as users tend to look at the interface objects back and forth several times if the interface layout is ambiguous or vague).

Case #2: Encountering difficulties in task performance

We also found that the degree of omission was affected by subjects' interaction with the tasks. When subjects said that they had difficulty finding a solution or were confused by the task instruction, their verbalizations remained at a very abstract level. This finding is consistent with Branch's, who observed that the number of "dead ends" encountered by the users affects the amount of data generated during the think-after [3].

Here we pick another case, Subject A11, who verbalized at a very abstract level because s/he was apparently having difficulties solving the problem. Subject A11 had a 71% total omission rate in his/her verbal AOI sequence.

According to the verbal report, the subject was working on question 4 of the airplane task and was looking at the first and the third graph from left to right: "so I was confused which one was (the right graph)…which I was trying to take...I was really very, very hesitant on this one...." This verbalization was assigned the following AOI sequence code: 6CE66E8 (6 is the code for the fourth question; 8 for one of the answer choices; C for the first AOI graph on cost; and E for the third AOI graph on capacity).

In contrast, the subject's eye movement AOI sequence showed the following: 6RQRQS6RQR6KS676SRSQKSM 86Q6SKSROPN6R86RS876S789786RLKSMQRMLSR7. We found that for question 4, the subject was constantly looking at the AOI fine levels, namely the first and third graphs. K,M, and L represent three fine level AOIs in the first graph; S,R, and Q are three fine AOIs in the third graph; 6,7,8,9 are four fine AOIs for the questions and answer choices. Although the subject mentioned that s/he was looking at the first and the third graphs, his/her verbalizations remained general and did not mention the fine level AOIs that he/she looked at. The alignment of verbal and eye movement AOI sequences resulted in 56 omissions, which accounted for 50% of the subject's total omissions. It should be noted that a large number of omissions also occurred for question 3 on the same airplane task and for the same reason, that the subject found the task to be confusing and thus scanned the materials repeatedly.

This case appears to exemplify what occurs when subjects are struggling to work out a solution without too much success; subjects tend to heavily revisit information sites that show up in the ET data coding as one long AOI sequence. However, the brevity of the subjects' retrospective verbalizations belies what their eye movement is telling us and may explain our finding of the significant effect of task complexity on omission. It appears that when participants work on a complex task that they have difficulty solving, they tend to experience equal difficulty in formulating and articulating how they went about solving the problem. When that happens, their retrospective verbalizations tend to be abstract and unclear, and any details about what they were attending to are missing.

We see a similar problem in concurrent think aloud when subjects fall silent at the points where the cognitive load is heaviest. It appears at this preliminary stage that stimulated RTA may not help us address this problem. This result, combining with the valid account given from complex tasks, indicates that the retrospective think-aloud could be a useful method for finding usability problems (based on valid account), but maybe not be a useful method for revealing all of the detailed steps in task performance (because of omissions). This issue calls for more research.

The concept of task complexity in this study is a function of information load, information diversity, information transformation, task-solution dimensions, and task steps. All these factors are constitutive of the tasks that we designed. But the combination of factors makes it difficult to isolate the one factor or factors that make the task harder to complete and harder to articulate. We should emphasize, though, that task complexity does not necessarily result in a poorer task or verbal performance, which also depends on a person's prior knowledge and work experience. Further investigation on how task complexity and prior knowledge may affect a person's verbalization needs to be done. It does not, however, fall within the scope of this paper.

CONCLUSIONS AND FUTURE WORK

In this study, we empirically investigated the validity and reliability of stimulated retrospective think aloud (RTA). Our study supported the validity of stimulated RTA in that people's recounting of what went on in their task performance describes the same sequence of objects in the same order as what they attended to during the original task performance. Our study also shows that the validity of the RTA doesn't vary with different levels of task complexity. These findings are useful in any field that uses RTA to collect user's performance information.

This study also shows that the logic inference and strategy explanation information in people's verbalization also provide valid information about users' task performance. This inferential and explanatory information can indicate how information was processed and clarify what specific strategies people used to complete tasks in a usability study. Usability evaluators can use this information to assess whether a product or interface is successful in supporting users in doing the tasks it is designed for and to identify what parts of the design negatively affect user's behavior.

Two case analyses about omissions in the verbal report show that, in general, in instances when users were struggling to complete tasks, the verbal reports provide low density and high abstract level information. Such patterns could be used as a valid indication of problems in a usability study.

The results and findings presented in the paper are preliminary work to establish the fundamental validity of stimulated RTA. Future works can be done following two trends. One is to design an advanced algorithm to extract high level information from ET, so that it could be used to compare with high level verbalization. Second is to study how a specific dimension of task complexity affects the degree of omission found in people's retrospective verbalization.

ACKNOWLEDGMENTS

We would like to thank Sandra Maddox, Crystal Chin, Yang Li, and James Landay for their comments and help in designing and piloting this study.

REFERENCES

1. Bell, B., et al. Usability testing of a graphical programming system: things we missed in a programming walkthrough. In *Proc. CHI'91*. ACM Press (1991), 7-12.
2. Bowers, V.A.&H.L. Snyder. Concurrent versus Retrospective Verbal Protocol for Comparing Window Usability. In *Proc. of the Human Factors Society 34th Annual Meeting*. (1990), 1270-1274.
3. Branch, J.L. Investigating the Information-Seeking Processes of Adolescents: The Value of Using Think Alouds and Think Afters. *Library & Information Science Research*. *22*,4 (2000), 371-392.
4. Campbell, D.J. Task Complexity: A review and analysis. *The Academy of Management Review*. *13*,1 (1988), 40-52.
5. Capra, M.G. Contemporaneous versus Retrospective User-Reported Critical Incidents in Usability Evaluation. In *Proc. of Human Factors Society, 46 th Annual Meeting*. (2002), 1973-1977.
6. Card, S.K., et al. Information scent as a driver of web behavior graphs: results of a protocol analysis method for web usability. In *Proc. CHI'01*. ACM Press (2001), 498-505.
7. Choi, B., et al. A Qualitative Cross-National Study of Cultural Influences on Mobile Data Service Design. In *Proc. CHI 2005*. ACM Press (2005), 661-670.
8. Ericsson, K.A.&H.A. Simon, Protocol analysis: Verbal Reports as Data. 1993: Cambridge, MA: MIT Press.
9. Gapra, M.G. Comtemporaneous versus Retrospective User-reported Critical Incidents in Usability Evaluation. In *Proceedings of the Human Factors and Ergonomics Society, 46th Annual Meeting*. (2002), 1973-1977.
10. Geiselman, R.E.&F.S. Bellezza. Eye-movements and overt rehearsal in word recall. *Journal of Experimental Psychology: Human Learning and Memory*. *3*,3 (1977), 305-315.
11. Gero, J.S.&H.-h. Tang. Differences between retrospective and concurrent protocols in revealing the process-oriented aspects of the design process. *Design Studies*. *21*,3 (2001), 283-295.
12. Goldberg, J.H.&A.M. Wichansky, Eye tracking in usability evlauation: A practitioner's guide., in *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements*, R. Radach, et al., Editors.(2003), Elsevier Science: Oxford. 493-516.
13. Gray, W.D.&M.C. Salzman. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*. *13*,3 (1998), 203-261.
14. Kensing, F. Prompted Reflections: A Technique for Understanding complex work. *Interactions*. *Jan-Feb.*,(1998), 7-15.
15. Kjeldskov, J.&M.B. Skov. Creating Realistic Laboratory Settings: Comparative Studies of Three Think aloud Usability Evaluations of a Mobile System. In *Proc. of the 9th IFIP TC13 INTERACT 2003*. (2003), 663 – 670.
16. Mankoff, J., et al. Is Your Web Page Accessible? A Comparative Study of Methods for Assessing Web Page Accessibility for the Blind. In *Proc. of CHI'05*. ACM Press (2005), 41-50.
17. Nielson, J., Usability Engineering. 1993: Cambridge, MA: AP Professional.
18. Page, C.&M. Rahimi. Concurrent and Retrospective Verbal Protocols in Usability Testing: Is There Value Added In Collecting Both? In *Proc. of the Human Factors and Ergonomics Society, 39th Annual Meeting*. (1995), 223-227.
19. Preece, J., Human-Computer Interaction. 1994: Addison-Wesley, England.
20. Preece, J., et al., Interaction Design: Beyond Human-Computer Interaction. 2002: John Wiley & Sons.
21. Ramey, J., et al., Adaptation of an Ethnographic Method for Investigation the Task Domain in Diagnostic Radiology, in *A Field Methods Casebook for Software Design*, e. D. Wixon and J. Ramey, Editor.(1996), John Wiley and Sons. 1-15.
22. Rhenius, D.&G. Deffner. Evaluation of Concurrent Thinking Aloud using Eye-tracking Data. *Proc. of the Human Factors and Ergonomics Society 34th Annual Meeting*. (1990), 1265-1269.
23. Rowley, D.E. Usability Testing in the field: bringing the laboratory to the user. In *Proc. CHI'94*. ACM Press (1994), 252 - 257.
24. Russo, J.E., et al. The Validity of Verbal Protocols. *Memory and Cognition*. *17*,6 (1989), 759-769.
25. Sankoff, D.&J.B. Kruskal, An overview of sequence comparison, in *Time Warps, String Edits, and Macro-Molecules: The Theory and Practice of Sequence Comparison*.(1983), Addison-Wesley.
26. Soukoreff, R.W.&I.S. MacKenzie. Measuring errors in text entry tasks: An application of the Levenshtein string distance statistic. In *Proc. CHI'01*. ACM Press (2001), 319-320.
27. St. Amant, R.&M.O. Riedl. A perception/action substrate for cognitive modeling in HCI. *International Journal of Human-Computer Studies*. *55*,1 (2001), 15-39.
28. Suwa, M.&B. Tversky. What architects see in their sketches: implications for design tools. In *Proc. CHI'96*. ACM Press (1996), 191-192.
29. Teague, R., et al. Concurrent vs. Post-Task Usability Test Ratings. In *Proc. CHI'01*. ACM Press (2001), 289-290.
30. Van den Haak, M.J., et al. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour& Information Technology*. *22*,5 (2003), 339-351.
31. Waes, L.V. Thinking Aloud as a Method for Testing the Usability of Websites: The influence of Task Variation on the Evaluation of Hypertext. *IEEE Transactions on Professional Communication*. *43*,3 (2000), 279-291.
32. Williams, T.R., et al. Does Isolating a Visual Element Call Attention to It? Results of an Eye-tracking Investigation of the Effects of Isolation on Emphasis. *Technical Communication*. *52*,1 (2005), 21-26.
33. Wood, R.E. Task Complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*. *37*,(1986), 60-82.