Which Is the Better Prompt in Thinking-Aloud Studies, “What Are You Trying to Achieve?” or “Keep Talking”?

Morten Hertzum
Computer Science
Roskilde University
4000 Roskilde, Denmark
mhz@ruc.dk

Kristin D. Hansen
Computer Science
Roskilde University
4000 Roskilde, Denmark
kdh@ruc.dk

Hans H. K. Sønderstrup-Andersen
National Research Centre for
the Work Environment
2100 Copenhagen, Denmark
hsa@nrcwe.dk

ABSTRACT

Thinking aloud is widely used for usability evaluation but generally in a relaxed way that conflicts with the prescriptions of the classic model for obtaining valid verbalizations of thought processes. We investigate whether participants that think aloud in the classic or relaxed way behave differently compared to performing in silence. Results indicate that whereas classic thinking aloud has little or no effect on behaviour apart from prolonging tasks, relaxed thinking aloud affects behaviour in multiple ways. During relaxed thinking aloud participants took longer to solve tasks, spent a larger part of tasks on general distributed visual behaviour, issued more commands to navigate both within and between the pages of the web sites used in the experiment, and experienced higher mental workload.

Keywords

Thinking aloud, verbalization, usability evaluation

INTRODUCTION

Evaluation is an important and complex element of systems development, and effective and valid evaluation methods are, consequently, in high regard. The thinking-aloud method has become popular in practical usability evaluation as well as in usability research and is by many considered the single most valuable usability evaluation method [3, 13]. However, the possible effects of thinking aloud on the behaviour of participants in usability evaluations have only been examined in a few studies [e.g., 8, 11]. Instead, claims to validity have been adopted from studies in cognitive psychology, particularly Ericsson and Simon’s [5] work on verbal reports. Descriptions of the thinking-aloud method for usability evaluation differ, however, in important respects from thinking aloud as defined by Ericsson and Simon [5, 6], particularly regarding instructions and reminders to think aloud, and these differences are likely exacerbated in practical use of the method [1, 14].

This study aims to investigate whether thinking aloud causes participants in usability evaluations to behave differently and experience a different level of mental workload compared to performing in silence. To address the variation in – and uncertainty about – what test participants are more specifically asked to do when they are asked to think out loud we distinguish between classic thinking aloud and relaxed thinking aloud. Classic thinking aloud complies with the prescriptions of Ericsson and Simon [5]. Relaxed thinking aloud complies with typical descriptions of thinking aloud in the context of usability evaluation [e.g., 4, 13]. We restrict our study to concurrent thinking aloud; that is, participants think aloud while solving tasks.

Usability evaluation has become widely practiced, not least through the uptake of lightweight, or discount, methods [e.g., 12, 13]. These methods promise to require little time, few skills, hardly any facilities, and yet to yield good results. The entire approach stands in stark contrast to the rigour of Ericsson and Simon’s [5] description of thinking aloud and to their assessment of the consequences of failing to think aloud in the proper way. Ericsson and Simon [5] distinguish three levels of decreasingly valid verbalizations, each characterized by the amount of interference caused by the additional processing involved in producing the verbalizations:

Level 1 verbalization is the vocalization of thoughts and information that are already in a person’s present focus of attention in verbal form. No intermediate processes are needed to report these thoughts and people need expend no special effort to communicate them. For example, when people report sequences of numbers while mentally solving math problems they are producing level 1 verbalizations because the reported numbers – that is, the intermediate results of the calculations – are directly available in the form needed to report them.

Level 2 verbalization is the explication of information that is presently in a person’s focus of attention but must be recoded into verbal form before it can be reported. The explication or recoding involves additional processing but does not bring new information into the person’s focus of attention. For example, images and abstract concepts must be transformed into words before they can be reported but

In O.W. Bertelsen and A.M. Kanstrup (eds.), *DHRS2009: Proceedings of the Ninth Danish HCI Research Symposium* (Aarhus, DK, December 14, 2009), pp. 30-33. DAIMI Report PB-591, Aarhus University, Aarhus, DK.

as long this transformation is the only additional processing that is performed such verbalization is at level 2.

Level 3 verbalization introduces mental processing that influences a person's focus of attention in ways beyond those occasioned by task performance. The influence on the person's focus of attention consists of requiring that present thoughts and information attended to at the moment are linked to earlier thoughts and information attended to previously. People, for example, produce level 3 verbalizations when they are asked to provide explanations of their thoughts and behaviour or to retrieve additional information from memory.

According to Ericsson and Simon [5] verbalizations at levels 1 and 2 are valid data about task performance, whereas level 3 verbalizations are not. This restricts valid verbalization to the currently heeded information; that is, the contents of short-term memory.

METHOD

Participants

Eight participants took part in the experiment. Participants' age ranged from 23 to 33 years with an average of 28.5 years. All participants were experienced computer users and indicated that they used computers daily. In addition, the participants had normal or corrected-to-normal vision and none used hard contact lenses or multi-focal glasses.

Tasks

The experimental tasks involved looking for information on four web sites – two web sites for Danish television channels and two for online bookstores. Each task was paired with another, near-identical task. The two tasks in a pair were performed on the web sites for either the two television channels or the two online bookstores. That is, the tasks in a pair were performed on similar but different web sites, thus reducing any learning effects. To further even out effects of learning, the order in which participants solved the tasks in a pair was counterbalanced across participants. Each participant performed 8 fact tasks and 8 assessment tasks. In *fact* tasks participants gathered information that was explicitly available on the web sites. For example, 'Which city has the highest temperature today – Copenhagen or Aarhus?' In *assessment* tasks participants gathered information and based on this information formed an opinion. For example, 'What is the biggest domestic news story on the front page?'

Procedure

Upon arriving at the lab, participants were introduced to the experiment and asked questions about their background. Then, participants were instructed about how to think aloud at levels 1 and 2 and practiced thinking aloud on four training tasks. The thinking-aloud instructions were copied from Ericsson and Simon [5, pp 377-379], and the three first training tasks were near identical to their training tasks. After practicing thinking aloud, participants were introduced to the task load index (TLX [9]). The preparations for the experimental tasks were completed by setting up and calibrating the eye tracker so that it accurately captured the participant's line of gaze. Participants' eye movements were recorded with a head-mounted eye tracker from SMI, sampling at 50 Hz.

The experiment consisted of two sub sessions for each participant. In the first sub session participants performed tasks in the classic thinking aloud and silent conditions. In *classic thinking aloud* participants performed the tasks while thinking out loud and the experimenter reminded participants to 'keep talking' when they fell silent for more than about 30 seconds. This condition corresponds to how thinking aloud is defined by Ericsson and Simon [5] as consisting of verbalization at levels 1 and 2. In the *silent* condition participants performed the tasks without verbalizing their thoughts. Participants were simply instructed to solve the tasks and report their answer to the experimenter upon completion. This condition is similar to how people work when they are not enrolled in usability evaluations or other tests. The two thinking-aloud conditions in the second sub session were relaxed thinking aloud and silent. In *relaxed thinking aloud* participants performed the tasks while thinking out loud and the experimenter intervened with questions asking participants for explanations and comments. Examples of the questions include 'What are you trying to achieve?' and 'Did you find this easy or difficult?' This condition includes level 3 verbalization and corresponds to how thinking aloud is commonly employed in the context of usability evaluation. The *silent* condition was similar to the first sub session.

RESULTS AND DISCUSSION

We investigated the presence of effects with respect to correctness of task solutions, task completion times, eye movements, hand movements, and mental workload.

The *correctness of task solutions* was not affected by whether participants were thinking aloud or performing in silence. This was the case for both classic and relaxed thinking aloud (but was determined for fact tasks only). For relaxed thinking aloud, some previous studies [2, 16] indicate that verbalization leads to fewer errors, compared to performing in silence.

Task completion times were longer during thinking aloud than when participants performed in silence. This difference was present for classic thinking aloud as well as for relaxed thinking aloud, and it was mainly due to participants' performance on assessment tasks, see Tables 1 and 2. The extra time required during thinking aloud accords with previous studies [5, 15] and indicates that verbalization is a slower process than thinking.

Participants' *eye movements* differed in some respects but did not provide evidence of a trend indicating that marginal effects for classic thinking aloud become significant for relaxed thinking aloud. During classic thinking aloud mental activity may have been shifted slightly from the start toward the end of assessment tasks. A similar effect was not found for relaxed thinking aloud. During relaxed

Table 1. Task completion times, *classic* thinking aloud

Task completion time		Classic		Silent	
(seconds/task)		*M*	*SD*	*M*	*SD*
Fact tasks		110	47	82	35
Assessment tasks	*	303	92	217	41

M – mean, *SD* – standard deviation, * $p < 0.05$

Table 2. Task completion times, *relaxed* thinking aloud

Task completion time		Relaxed		Silent	
(seconds/task)		*M*	*SD*	*M*	*SD*
Fact tasks		201	55	131	76
Assessment tasks	**	319	148	114	49

M – mean, *SD* – standard deviation, ** $p < 0.01$

thinking aloud saccades were of shorter duration for assessment tasks, the more complex type of tasks. This is often seen as an indication of decreased visual search [7]. However, in this condition participants also spent a larger part of task completion times in general distributed visual behaviour, see Table 3. This seems to indicate that participants to a larger extent needed to fixate briefly on various screen elements to assess their relevance and contribution to the tasks. Distributed visual behaviour is akin to visual search but at a level of aggregation where brief fixations intersperse an activity primarily characterized by frequent saccades between screen elements that are spatially far apart and typically also distinct in contents. One reason for the increase in general distributed visual behaviour during relaxed thinking aloud could be that verbalizing at level 3 disrupted participants' mental activities and made it more difficult for them to maintain a focus, necessitating more distributed visual behaviour to regain a focus. Another reason could be that relaxed thinking aloud made participants in doubt about their approach to solving tasks or aware of other ways of solving them, leading to more distributed visual exploration of the screen.

Table 3. Visual behaviour, *relaxed* thinking aloud

Eye movements		Relaxed		Silent	
(percent of task time)		*M*	*SD*	*M*	*SD*
Focused: general		23	8	21	7
Focused: text		3	4	5	7
Focused: illustration		3	3	2	2
Distributed: general	**	10	4	5	5
Distributed: text		38	7	39	17
Distributed: illustration		1	1	1	2

M – mean, *SD* – standard deviation, ** $p < 0.05$

Table 4. Hand movements, *relaxed* thinking aloud

Hand movements		Relaxed		Silent	
(pr task)		*M*	*SD*	*M*	*SD*
Mouse clicks	*	5.9	4.9	4.0	2.6
Scrolling instances	**	26.6	21.5	10.9	8.9
Writing instances		2.4	1.8	1.9	1.0

M – mean, *SD* – std deviation, * $p < 0.05$, ** $p < 0.01$

Participants' *hand movements* revealed considerable differences in how participants interacted with the system while solving the tasks. Mouse clicks and writing instances approached significantly higher numbers during classic thinking aloud compared to the silent condition. This suggests that participants may be paying more attention to obtaining information from web pages other than the current one. During relaxed thinking aloud participants made more mouse clicks and scrolling instances compared to the silent condition, see Table 4. This indicates that participants made more efforts to obtain information from other web pages by making more shifts between web pages, and they also explored the current web page more comprehensively by scrolling more. Hence, the increase in task completion times must be interpreted differently for classic and relaxed thinking aloud. During classic thinking

Table 5. Mental workload, *classic* thinking aloud

TLX subscale		Classic		Silent	
(0-100)		*M*	*SD*	*M*	*SD*
Mental demand	*	41	26	31	17
Physical demand		20	22	14	13
Temporal demand		23	22	20	13
Effort		28	24	21	14
Performance		30	26	29	23
Frustration		26	19	18	13

M – mean, *SD* – standard deviation, * $p < 0.05$

Table 6. Mental workload, *relaxed* thinking aloud

TLX subscale		Relaxed		Silent	
(0-100)		*M*	*SD*	*M*	*SD*
Mental demand	**	30	15	19	10
Physical demand		16	13	10	7
Temporal demand	*	18	12	10	7
Effort	*	25	12	14	8
Performance	*	21	18	17	16
Frustration	*	21	15	10	7

M – mean, *SD* – std deviation, * $p < 0.05$, ** $p < 0.01$

aloud the increase in task completion times was primarily a slowdown in participants' performance but during relaxed thinking aloud participants performed the tasks in a different way.

Mental workload approached higher ratings for classic thinking aloud, compared to performing in silence. For relaxed thinking aloud participants rated mental workload higher than for performing in silence. This overall picture was repeated for the individual TLX subscales. Verbalization at levels 1 and 2 added to one of the six subscales of mental workload, whereas verbalization at level 3 added to all but one of the subscales, see Tables 5 and 6. The results for mental workload, a subjective measure, are consistent with the performance measures. Effect sizes tend to be larger for mental workload than for the performance measures, suggesting that participants may moderate performance differences by putting in extra mental effort while thinking aloud.

CONCLUSION

Thinking aloud is widely used as a method for usability evaluation. The method is, however, generally applied in a relaxed manner that conflicts with the prescriptions of Ericsson and Simon's classic model for obtaining valid verbalizations of thought processes. Descriptions of thinking aloud in the methodological literature often display a similar failure to consistently distinguish between classic thinking aloud (corresponding to verbalization at levels 1 and 2) and relaxed thinking aloud (corresponding to verbalization at levels 1 to 3). In this study, we have investigated the effects of thinking aloud over performing in silence for both classic and relaxed thinking aloud.

Our results confirm that classic thinking aloud has little effect on participants' behaviour and mental workload, except for prolonging tasks. Hence, valid data about the use of a system can be obtained at the price of precise instructions and minimal interaction between the user who thinks aloud and the evaluator who listens in on the user's thoughts. Conversely, relaxed thinking aloud led to longer task completion times, a larger part of tasks spent on general distributed visual behaviour, more commands issued to navigate both within and between the pages of the web sites used for solving the tasks, and higher perceived mental workload. Hence, the relaxed approach to thinking aloud threatens the validity of the method and indicates that this approach, common in practical usability evaluation, may not be the authoritative yardstick it is often assumed to be.

ACKNOWLEDGEMENTS

This paper is a summary of Hertzum et al. [10]. Please, consult the full paper for the details.

REFERENCES

1. Boren, T., and Ramey, J. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication 43*, 3 (2000), 261-278.
2. Chi, M.T.H., De Leeuw, N., Chiu, M.-H., and LaVancher, C. Eliciting self-explanations improves understanding. *Cognitive Science 18*, 3 (1994), 439-477.
3. Dumas, J.S. User-based evaluations, in Jacko, J. and Sears, A. eds. *The Human-Computer Interaction Handbook*, Erlbaum, Mahwah, NJ, 2003, 1093-1117.
4. Dumas, J.S., and Redish, J.C. *A practical guide to usability testing. Revised edition*. Intellect Books, Exeter, UK, 1999.
5. Ericsson, K.A., and Simon, H.A. *Protocol analysis: Verbal reports as data. Revised edition*. MIT Press, Cambridge, MA, 1993.
6. Ericsson, K.A., and Simon, H.A. Verbal reports as data. *Psychological Review 87*, 3 (1980), 215-251.
7. Goldberg, J.H., and Kotval, X.P. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics 24*, 6 (1999), 631-645.
8. Haak, M.J.v.d., Jong, M.D.T.d., and Schellens, P.J. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology 22*, 5 (2003), 339-351.
9. Hart, S.G., and Staveland, L.E. Development of NASA-TLX (task load index): Results of empirical and theoretical research, in Hancock, P.A. and Meshkati, N. eds. *Human Mental Workload*, North-Holland, Amsterdam, 1988, 139-183.
10. Hertzum, M., Hansen, K.D., and Andersen, H.H.K. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology 28*, 2 (2009), 165-181.
11. Krahmer, E., and Ummelen, N. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication 47*, 2 (2004), 105-117.
12. Monk, A., Wright, P., Haber, J., and Davenport, L. *Improving your human-computer interface: A practical technique*. Prentice Hall, New York, 1993.
13. Nielsen, J. *Usability engineering*. Academic Press, Boston, 1993.
14. Nørgaard, M., and Hornbæk, K. What do usability evaluators do in practice? An explorative study of think-aloud testing, in *Proceedings of the Sixth DIS Conference on Designing Interactive Systems*, ACM Press, New York, 2006, 209-218.
15. Rhenius, D., and Deffner, G. Evaluation of concurrent thinking aloud using eye-tracking data, in *Proceedings of the Human Factors Society 34th Annual Meeting*, HFS, Santa Monica, CA, 1990, 1265-1269.
16. Wright, R.B., and Converse, S.A. Method bias and concurrent verbal protocol in software usability testing, in *Proceedings of the Human Factors Society 36th Annual Meeting*, HFS, Santa Monica, CA, 1992, 1220-1224.