MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-309

Cognitive Evaluation of a Physician Data Query Tool for a National ICU Registry: Comparing Two Think Aloud Variants and Their Application in Redesign

Linda W. Peute, Nicolette F. de Keizer, Monique W. M. Jaspers

Department of Medical Informatics, Academic Medical Center – University of Amsterdam, The Netherlands

Abstract

Applying usability methods in formative evaluations of interactive healthcare information systems design is recognized as of extreme importance to the final success of these systems. However, it seems that the merits of specific methodological approaches for conducting these studies have received little attention. This study reports on a cognitive evaluation of a Physician Data Query Tool, which offers physicians the opportunity to query quality of care data collected by the Dutch National Intensive Care Evaluation (NICE) foundation. A comparison in terms of usefulness and utility of two variants of the Think Aloud method is addressed, the Concurrent and Retrospective Think Aloud. These methods are well known in the field of Human Computer Interaction in the context of usability evaluation. The results of this research indicate that though both methods have their disadvantages and benefits, in redesigning the Physician Data Query tool the Retrospective Think Aloud provided more useful input to the Tool's redesign. However, in deciding which method to apply in a formative evaluation study, end users' cognitive workload of performing the system's tasks and the system characteristics need to be considered as well.

Keywords:

User-Computer Interface, Usability Study, Cognitive Evaluation, Health Care Quality Evaluation, Physician Data Query.

Introduction

Understanding the cognitive processes of clinicians, such as the processes by which they (learn to) comprehend and interact with interactive programs, is a prerequisite for building tools that support clinical practice in an appropriate manner [1]. The think aloud method is generally recognized as a major source of data on subjects' cognitive processes and has been applied in studies on computer program comprehension in the field of Human-Computer Interaction (HCI) for decades. Its application in usability testing of health information systems aims to improve clinician system interaction and to develop more usable interfaces [2]. Ericsson and Simon (1984) have presented two variants of the think aloud methods which have been applied in the context of usability testing in recent years; Concurrent (CTA) and Retrospective (RTA) Think Aloud [3,4]. When applying the CTA method, subjects are instructed to verbalize their thoughts while conducting a task. In contrast, the RTA method instructs users to recall their thoughts or actions after they have finished the task by for example reviewing a video recording of their actions. Analysis of the verbal protocols and video recordings has the purpose of revealing the contents of the subject's working memory and his corresponding action, thus providing a unique insight into the subject's cognitive processes in relation to experienced system usability. Controversy exists, however, about the utility and validity of these two methods in usability testing of interactive health information systems [4,5]. It seems that the choice of application of these methods should be based on the nature and complexity of the task to be performed by the subject [4]. Also, which method best to apply in a formative evaluation study of an interactive health information system is still unclear.

This paper explores the use and utility of both the Concurrent (CTA) and Retrospective (RTA) Think Aloud method in a usability study of a Physician Data Query Tool for a national ICU quality of care registry in the Netherlands. We address a comparison of the methods in task completion time and task performance with tasks of differing cognitive difficulty, and type and number of usability problems detected. We discuss the implication of these results for system redesign in general and in light of the efforts that are currently undertaken in redesigning the Physician Data Query Tool.

Materials and Methods

Test object: NICE Online, a Physician Data Query Tool

In 1996 the Dutch National Intensive Care Evaluation (NICE) foundation started collecting data on patients admitted to Dutch ICUs. The NICE database, also called NICE registry, contains information on demographic, physiological and clinical variables required to calculate mortality risk predictions according to the Intensive Care prognostic models [6]. The NICE registry aims to detect differences and trends in quality and efficiency of ICU care and provides quality reports and benchmarking information to its participating hospitals on a quarterly basis. In 2004 the request was made by participating hospitals if the NICE database could be queried by them to answer more specific clinical questions for their own ICU management or scientific reporting. To provide participants with the opportunity to query the NICE database

while protecting the privacy of the participating hospitals, a Physician Data Query Tool called **NICE Online** was developed in 2004 by a group of software engineers.

A standard software design cycle was applied in this project with the primary focus to develop a graphical interface for querying purposes. The user's view on query commands was reduced to a designer-customized browser, with a structured query model (figure 1) to support clinicians in developing queries, assuming that they have no or little experience in database query command development.

Figure 1-Screenshot of the Physician Clinical Data Query Tool: NICE Online.

Nice Online can only be used by participants with a user account. When entering the system, a large list of 'standard queries' is presented to the user. He/she can decide to select one of these queries and directly view the query's resulting table or graph or he/she can chose to change the query by adding (data) elements to the query model. Another possibility is to start a new query, also called 'custom query' in the system. A user is then presented with a blank query model, in which he/she must add all (data) elements needed for the query. The query model consists of four components: functions, splitting/intersection, benchmarking/mirror and selection of subpopulation. For each component the user can select from a large list of elements that are either statistical models or data elements collected by the NICE Registry. For example, functions are Intensive Care prognostic models which may form the basis of a clinical query. 'Benchmarking' refers to at least the user's 'own hospital data' for example in comparison to the element 'national data'. The splitting/intersection and selection of subpopulation components offer the user the possibility to split the data for example in gender categories or to create a subpopulation with regard to for example a certain time period. When a user is finished with the query model, he/she can select the 'graph/table' button to create the resulting graph or table.

Not all participants have yet requested a user account of NICE Online. In July 2008, NICE Online registered 80 users. A log file analysis, performed to gain insight into NICE Online usage patterns, showed that only 17% of the participants with a user account actually used the query functionality on a regular basis. Telephonic information needs analysis provided insight into user experiences with NICE Online. It showed that users were willing to use the tool but the structured interface for query development was not appreciated by them. However it did not become clear in what way the cognitive burden of a query development by use of the Tool was influenced by a potential lack in the Tool's usability. Next to this, planned expansions to the NICE Database, such as the collection and reporting of structure, process and outcome quality indicators of ICU's, required a high level of user-friendliness of NICE Online, which made it necessary to redesign the Tool and improve on its usability.

Subjects

Our study involved 16 subjects. Subjects were categorized on basis of a log file analysis of NICE Online usage patterns in order to select a number of representative target users. Subjects were assigned to one of the two conditions in a matched-randomized way with tool experience as matching factor. For testing of the Physician Data Query Tool formal agreement to contact the users was given by the NICE Foundation. The selected subjects were then contacted by email with an accompanying letter. All agreed to participate to the NICE Online evaluation study.

Tasks

To evaluate the cognitive workload and the usability of the Physician Data Query Tool (NICE Online), six tasks were developed which were divided into two to six smaller subtasks. Input into the development of these tasks was given by two data managers of the NICE Foundation. They where highly experienced in ICU clinical query development and were able to provide generally relevant tasks of varying difficulty with a golden standard for how to perform and finalize each task. The tasks were preceded by a query description, or short clinical question, which could be answered by data in NICE Online. The usability test started with two standard query tasks, randomly given. These tasks provided the subject with some experience in NICE Online, and for the CTA method it provided participants with practice in verbalizing their thoughts while performing a task in the system. Then four tasks were randomly presented to the subject with a varying degree of difficulty. These four tasks consisted of two custom query tasks, in which the user had to enter a query statement as described in NICE Online, and two tasks consisting of a clinical question to be translated into a query in NICE Online. An example of both tasks is given in Table 1.

Table 1- Examples of the usability Tasks

	Examples Main question
Custom query task *easy*	*'Please select the percentage of patient admissions which are split by admission type for the data of your own hospital within a sub selection of the last two years'*
Translating a clinical question task *easy*	*'The annual NICE report shows you that there exists a difference in the mean length of stay for patients in the age category 60 to 70 year in the year 2007 compared to 2008. You whish to find out if this is correct by making a graph of these data in NICE Online.'*

CTA and RTA experiment

The experiments took place in the actual clinical working area of the subjects. A portable usability laptop with Morae software made it possible to record all the subjects' verbalizations in combination with a screen recording of their (mouse) actions in the system and a video recording of the subjects performing the actual tasks in NICE Online on the usability laptop. In both the CTA and RTA condition, the experimental procedure started with the subject answering questions about his or her general computer experience, experience with statistical knowledge and calculations, and experience in formulating queries in other systems. Hereafter the subject received the tasks as well as oral instructions on how to carry them out on the laptop. In the CTA condition, the subject was instructed to think aloud while performing the tasks. In line with the Think Aloud procedure described in [3], it was made clear that the accompanying researcher, the facilitator, would not interfere with the session by giving assistance, but would only remind the subject to keep thinking aloud if the subject would fall silent for a while. Finally, the subject was told that the goal of this test was to gain insight into the problems he/she might encounter in using NICE Online, and to understand in what way he/she translated a clinical question to a NICE Online query. In the RTA condition, the subjects received the tasks and short oral instructions. They were instructed to carry out the tasks in silence on the laptop, without assistance of the facilitator. After the session, video recordings of their actions in the system were shown to them and they were asked to verbalize their thoughts retrospectively. The analysis of the think-aloud sessions was done in the Morae Manager from Techsmith. Additional validation of the subjects' task performance was performed by the research and database manager of the NICE Foundation.

Results

General results

The 16 experiments resulted in over 24 hours of recordings. The CTA testing lasted approximately 1 hour, while the RTA testing lasted about 2 hours, including the time for retrospective reporting. The transcription of the verbal protocols of subjects in the CTA condition resulted in 3 times as much data compared to the transcription of the verbal protocols of subjects in the RTA condition. Of the subjects, 12,5% was female. Of all subjects 62,5% mentioned that they considered themselves expert with regard to computer experience, 50% considered themselves expert in statistical calculations and 56,25% regarded themselves expert in query development. The analysis showed that subjects who considered themselves as experts were somewhat equally divided between the two methods.

The verbal protocols of subjects were transcribed and all actions of subjects were linked to the comments made by them as reflected in the protocols. Two analysts went through all verbal protocols and video recordings and separately coded the usability problems, experienced by the subjects, in usability categories described among others by Kushniruk et al. [7]. Inter rater reliability was measured by Cohen's kappa (.83) which constitutes to a substantial agreement between the two analysts. In total 43 singular usability problems were analyzed.

Task completion and performance

Table 2 shows a comparison of the CTA and RTA method with respect to 1) task completion time in minutes and 2) task performance in terms of incorrect tasks (N/E) per difficulty level. Analysis showed that development of a custom query in the system took slightly more time in the CTA condition than in the RTA condition. Yet, there was only one clear difference in task completion time between the two methods. The translation of the clinical question to a query in NICE Online in the easy category took exceedingly more time in the CTA condition than the RTA condition. Also, this task was more often performed incorrectly by subjects in the CTA than in the RTA condition. Analysis of the verbal protocols revealed that subjects in both conditions experienced much difficulty in translating a clinical question to a query in the Tool and commented on this task to be cognitively complex. This translation proved however more difficult for subjects in the CTA condition than in the RTA condition. In total, 24 tasks were incorrectly performed in the CTA condition whereas 12 were incorrectly performed in the RTA condition. Overall, the CTA condition had lower task performance than the RTA condition.

Table 2 - Overall task completion time in minutes, N/E tasks not executed correctly

	CTA			RTA		
	Mean	**SD**	**N/E**	**Mean**	**SD**	**N/E**
Standard Query						
Easy	4.8	1.8	1	5.0	0.3	0
Difficult	7.3	1.8	2	6.9	2.0	0
Custom Query						
Easy	6.1	0.1	3	4.5	0.7	2
Difficult	7.5	2.0	4	6.8	2.1	3
Translating Question Query						
Easy	11.3	1.8	5	7.5	0.8	2
Difficult	11.7	0.9	9	9.4	3.3	5

Usability problems

Table 3 gives an overview of the mean number of usability problems detected for each usability category per think aloud method. Table 3 also shows the number of problems per category that was uniquely detected by one of the two methods (CTA or RTA) and those problems that were detected in both the CTA as well as the RTA condition. The CTA condition provided insight into several types of usability problems. Analysis of the verbal protocols revealed that subjects in the CTA condition, when confronted with usability problems of a minor or cosmetic nature that directly obstructed the performance of a task, directly commented upon that issue. In contrast, the verbal protocols of subjects in the RTA condition showed that subjects did not report upon

these minor usability problems. Instead, subjects' comments in the RTA condition focused more on the complex usability issues they had experienced during the test.

Table 3 - Usability problems per category and total number of problems per think aloud method.

Problem types	CTA		RTA		CTA	RTA	Both
	Mean	SD	Mean	SD	#	#	#
Navigation	4.9	1.2	4.2	1.1	2	0	3
Graphics/ symbols	5.6	1.8	3.4	0.5	4	2	2
Layout/screen organization	4.8	1.4	2.9	0.8	2	0	4
Meaning of labels/ terminology	5.0	0.5	7.2	1.2	0	4	4
Error messages/ help instructions	4.9	1.7	4.7	1.4	2	0	3
Overall ease of use	6.0	2.8	4.1	1.2	1	1	6
Visibility of system status	3.8	1.5	3.2	1.1	1	0	2
Total number of usability problems	36.0	-	31.0	-	12	7	24

Overall, it can be stated that the CTA condition revealed more usability problems than the RTA condition. The RTA method however provided more usability issues concerning the terminology and meaning of labels. Also, the verbal protocols of subjects in the RTA condition, subjects' verbalizations proved more explanatory towards these problems. For example, one subject in the CTA condition did not completely understand what the term splitting/intersection meant in the query model, he showed irritation and commented there upon, while the subject in the RTA condition explicitly described what the problem was, and how to resolve this terminology issue, which was of high importance to redesign of NICE Online. The analysis of the CTA and RTA verbal protocol data showed that subjects' verbalizations in the CTA and RTA method differed considerably. The golden standard for task completion provided by the NICE data managers proved useful in analyzing the verbal protocols of subjects. When tasks were commented upon by subjects in one of the two conditions, the numbers of statements made by a subject was classified in terms of 'experienced Tool usability', 'explanatory Tool usability', 'task statistical reasoning', 'task comprehension', and 'task query complexity' and were subsequently counted. The verbal protocols showed that comments made by subjects' about 'statistical reasoning', 'comprehension of the task' to be performed, or comments made about the 'complexity of the query' were of a different cognitive nature then the comments made on the experienced Tool usability. Table 4 shows the mean number of these 'cognitive' problems detected per method. In the CTA condition during the usability test subjects explicitly verbalized when they did not fully comprehend a statistical model or the task to be performed or the query to be made in the Tool. Subjects in the RTA condition did not comment upon a potential lack in their statistical knowledge, or in their (in)comprehension of the task at hand.

Table 4 – Mean number of statements of a cognitive nature per method

	CTA		RTA	
	Mean	**SD**	**Mean**	**SD**
Problems in statistical reasoning	5.3	1.0	1.0	1.1
Problems in task comprehension	5.1	1.7	2.3	1.2
Problems in query design complexity	5.1	1.4	4.5	2.2

Discussion

This study shows that task completion in the CTA condition for standard and custom query tasks did not take more time than task completion for standard and custom query tasks in the RTA condition. The task of translating a clinical question to a query in NICE Online took generally more time in both methods, but took exceedingly longer in the CTA condition than the RTA condition. In the Human Computer Interaction literature it is under debate if the CTA and RTA methods offer similar results in terms of task completion time [8,9]. The results of this study seem to indicate that subject's task completion time in the CTA method is influenced by the task complexity. Apparently task completion time of the translation of a clinical scenario to a query seemed to be affected by the cognitive workload of both query translation and direct verbalization of the corresponding actions in the system. Also, the number of tasks completed correctly was lower in the CTA than in the RTA condition. This might point out that the double workload of verbalizing thoughts and performing actions in the CTA method causes subjects' to make more errors or less easily recover from usability problems experienced in performing a task in the Tool. The cognitive complexity of translating a scenario to a query might affect subjects' task performance

In discussing the type of problems detected per method, it becomes clear that each method revealed unique usability problems. The CTA uncovered more usability issues in general, and more specifically it revealed problems concerning system graphics, navigation, error messages and the layout and organization of the computer screen. The CTA method also seemed to uncover more usability issues of a more cosmetic nature than the RTA method. In comparison, the RTA method uncovered more usability problems related to terminology and meaning of labels, and uncovered more usability issues of a complex nature. Also, its verbal protocols provided more explanatory verbalizations which were considered useful by the software engineers as they provided

information on how to resolve such usability issues in redesign of the NICE Online tool.

Another interesting finding of the RTA method in contrast to the CTA method was the fact that subjects did not verbalize potential problems they had experienced related to statistical, task and query comprehension. Therefore, a slightly positive bias towards the NICE Online tool could be seen in the RTA condition, suggesting that subjects made their comments more positive, thus disguising their 'not so good' result caused by a lack in statistical knowledge, or their possible incomprehension of the task to be performed. The subjects involved in this study were clinicians. A reason for this 'disguising behaviour' might be found in that a 'hospital culture' could exist which prevents them to comment on errors made following their lack in knowledge on for example clinical statistics [10]. However, if the RTA subjects indeed did not express their experienced difficulties and their need for more computer support in certain phases of task execution, then the results of the usability study may be less valid, and could for example lead to missing important insights into additional functionalities required in a system redesign. For example, NICE online is specifically designed to provide its users the opportunity of analyzing ICU data, also from a statistical point of view. While subjects in the CTA condition expressed their need for additional support in statistical reasoning to adequately make use of the Tool in the CTA condition, this requirement was not expressed by subjects in the RTA condition. Based on the verbalizations of the subjects in the CTA condition, the conclusion was drawn to provide additional functionality in the help and information support of the NICE Online Tool.

Our study not only showed that the cognitive difficulty of tasks influenced the total completion time, but also that subjects' task performance in terms of correct tasks was lower in the CTA condition than in the RTA condition. In this case, subjects were not supported by the Tool in adequately performing the task of translating a clinical question to a query, mainly because of the many usability problems that they surfaced. However, for NICE Online, not all of these usability problems might prove to be of importance in redesigning the cognitive model of clinical querying.

Conclusion

Overall, the conclusion can be drawn that the RTA method provided more useful information for the Tool's redesign, because of the more explanatory nature of the verbal protocols, and the provided insight into complex usability issues, Subjects had more time to articulate why they were having problems and they did not focus on the irritations caused by the experienced usability problems in performing their tasks. While the CTA method provided a useful overview of a large number of usability problems experienced including minor usability issues, some of its resulting verbalizations did not provide enough detail to support the Tool's redesign. Also, it is not clear if all these problems in redesign need to be coped with. However, these results also indicate that the RTA method might lead to missing input on additional functionalities in system redesign, because statements concerning physicians lack in task and statistical comprehension were nor adequately made. Though CTA has as its benefit that its subject testing takes less time than RTA, it is of importance to decide which method best to apply in revealing the usability issues that provide enough insight needed for a system's redesign. In light of these formative results, the Physician Data Query Tool is currently under redesign. A future study will focus on a comparison of these two methods in a pre-post design in usability testing of the redesigned Tool.

Acknowledgments

The authors would like to thank the NICE Foundation for making this study possible. Also all NICE participants who acted as subjects in this study are thanked for their contribution. Last but not least, thanks goes out to the head software engineer of NICE Online for his continuous support in conducting this study.

References

[1] Patel V, Kushniruk AW. Interface Design for Health Care Environments: The role of Cognitive Science. Proceedings AMIA Symposium 1998:121-5

[2] Horsky J, Zhang J, Patel VL. To err is not entirely human: complex technology and user cognition. Journal of biomedical informatics 2005: 38(4): 264-266.

[3] Ericsson KA, Simon HA. Protocol analysis: verbal protocols as data. MIT Press, Cambridge 1993.

[4] Bowers VA, Snyder HL. Concurrent versus Retrospective Verbal Protocol for Comparing Window Usability. Proceedings of the Human Factors Society 34th Annual Meeting 1990, 1270-1274.

[5] Jaspers MWM. A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. International Journal of Medical Informatics 2009: 78: 340-353.

[6] Website: http://www.stichting-nice.nl/index.jsp

[7] Kushniruk AW, patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. Journal of biomedical informatics 2004: 37: 56-76.

[8] Gray WD, Salzman MC. Damaged merchandise? A review of experiments that compare usability evaluation methods. Human Computer Interaction 1998: 13: 203-261.

[9] Molich R, Thomson AD, Karyukina B, Schmidt L, Ede M, Oei van W, Arcuri M. Comparative evaluation of usability tests 1999: http://www.dialogdesign.dk/cue.html.

[10]Kingston MJ, Evans SM, Smith BJ, Berry JG. Attitudes of doctors and nurses towards incident reporting: a qualitative analysis. The Medical journal of Australia 2004:181(1):36-39.

Address for correspondence

Department of Medical Informatics, Academic Medical Center, PO Box 22700, 1105 AZ Amsterdam, The Netherlands, E-mail: l.w.peute@amc.uva.nl