Human factors methods in health information systems' design and evaluation: The road to success?
Peute, L.W.P.

Citation for published version (APA):
Peute, L. W. P. (2013). Human factors methods in health information systems’ design and evaluation: The road to success?
Chapter 4

Effectiveness of Retrospective and Concurrent Think Aloud in Formative Usability Testing; Which Method Performs Best?

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

1. *This submitted study expands on a published Second Best Conference Paper Awarded and presented at MEDINFO 2010 - CapeTown.*

Abstract

Objective: To compare the performance of two usability testing methods, the Concurrent (CTA) and Retrospective (RTA) Think Aloud, in a formative usability study of an Intensive Care Registry Physician Data Query Tool.

Methods: Sixteen representative target intensivists participated in the usability evaluation study. Subjects were allocated to one of the method CTA/ RTA condition by a matched randomized design. Each subject performed six usability testing tasks of varying query development complexity in the Query Tool. The methods’ performance was compared in terms of overall effectiveness in detecting usability problems within the average time subjects spent per method. Subsequently, the thoroughness of the methods in detecting usability problems weighted by problem severity level and problem type was analyzed.

Results: The usability evaluation of the Data Query tool revealed a total of 43 unique usability problems. CTA performed significantly better than RTA in detecting usability problems. Overall, CTA usability problem detection effectiveness was 0.80 vs. 0.62 (p<0.05) respectively with an average difference of 42% less time spent per subject compared to RTA. In addition CTA was more thorough in detecting usability problems of a moderate (0.85 vs. 0.7) and severe nature (0.71 vs. 0.57). Regarding problem type detection, CTA detected unique usability problems concerning graphics/symbols, navigation issues, error messages and the organization of information on the Query Tool’s screens. RTA detected unique issues concerning system match with users’ language and applied terminology. Qualitative analysis of the CTA and RTA verbal protocols showed that RTA verbal protocols contained significantly more explanatory comments regarding the cause of a usability problem and comments concerning additional system requirements.

Conclusion: CTA is more effective in usability problem detection but does not outperform RTA. RTA additionally provides sight on unique usability problems and new user requirements for specific user groups. Based on the results of this study we recommend the use of CTA in formative usability evaluation studies of health information technology. However, we recommend further research on the use of RTA in usability studies focusing on user profile customized (re)design.

Keywords: Usability Evaluation, Think Aloud, Intensive Care Information System
Introduction

In the current era, clinicians are increasingly becoming dependent on interactive healthcare applications to provide them access to the information they require [1-3]. Easy navigation and high understandability of the application’s interface have therefore become imperative to clinicians for efficient and effective system use [4]. In other words, interactive healthcare applications need to be designed with explicit regard to their usability; where usability is defined as the effectiveness, efficiency and satisfaction with which specific users can achieve a specific set of tasks in a particular environment [5].

Formative usability studies provide means to improve on a system design by uncovering those interface design flaws that clinicians might encounter when interacting with a system in a clinical setting [6]. To perform a formative usability study a broad spectrum of Usability Evaluation Methods (UEMs) is available and these methods are increasingly used in interactive healthcare applications design and evaluation [7-8]. However, selection of a UEM in a specific healthcare setting is often limited by practicality and by accessibility of required human resources and time to perform the evaluation study. The spare number of publications on UEM performance in formative studies on interactive healthcare applications furthermore limits usability practitioners in motivated selection of a UEM [9]. UEMs detection scopes may however differ significantly and may prove to be more or less informative to system (re)design in different study settings. To support justified UEM selection this study compares the performance of two known UEMs; the Concurrent Think Aloud (CTA) and Retrospective Think Aloud (RTA), in a formative usability study of a web-based Intensive Care (ICU) Registry Data Query Tool.

CTA finds its foundation in the cognitive and behavioral sciences [13-14]. Its application in usability testing studies yet has shown its value in providing rich and valid data on the usability of a system’s design [15]. In CTA usability testing subjects are instructed to verbalize their thoughts while conducting predefined tasks in the system concurrently. The RTA is a variation on the CTA and emerged to bypass certain limitations of the CTA in usability testing studies [16]. Concurrent verbalization for instance might interfere or slow down subjects’ task performance and in doing so may influence the reliability of CTA usability testing measures such as task efficiency and efficacy [17]. In contrast to CTA, the RTA method instructs users to recall their thoughts and actions after they have finished the computer supported predefined task(s). So, during RTA subjects verbalize their thoughts while reviewing a screen recording of their performance while interacting with the system under study. In this way, no direct interference of a subjects’ task performance occurs. However, linguistic studies and studies from psychology have shown that, when compared, the methods do appear to collect verbalized data of a different quantity and differing kind [18,19]. A basic assumption for this difference in methods’ output can be found in the workings of the human memory. In
concurrent verbalizations thought processes are expressed during task performance while retrospective verbalizations rely on the retrieval (recall) of these thought processes on tasks already completed. However, if these methods differ in their collection of data they might also differ in their detection scope of usability problems. It is of importance to understand the methods’ differences in performance given that (re)design efforts of interactive health application are primarily based on a usability method’s results.

In this paper the following research questions are addressed: Are the CTA and RTA method comparable in terms of:

*Overall effectiveness* in detecting usability problems with regard to *time subjects spent* per method?

*Thoroughness* in detecting usability problems of a minor, moderate or high severity?

*Types* of usability problem detected?

---

**Background Test Object; ICU Data Query tool**

In 1996 the Dutch National Intensive Care Evaluation (NICE) foundation started a registry collecting data on patients admitted to Dutch Intensive Care Units (ICUs). A data set of about 100 items on demographic, physiological and clinical variables is collected for each individual patient. Based upon this data several prediction models such as APACHE can be calculated to correct measures of crude hospital mortality for illness severity at admission [20]. The 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 ICUs to develop a tool to support ICU management and scientific reporting. To accommodate their request, a web-based application for clinical data querying was developed called ‘NICE Online’. NICE Online provided NICE subjects the opportunity to query the NICE database while protecting the privacy of patients and hospitals included in the Registry.

A standard software design cycle was applied in 2005 in the development project of NICE Online. A graphical interface was build providing the functionality (utility) of querying collected ICU data and the utility to benchmark the ICU data collected in the NICE registry. The user’s view on query commands was reduced to a custom designed webpage showing a structured query model (figure 1) to support clinicians in developing queries themselves.
Figure 1. Screenshot of the Physician Clinical Data Query Tool: NICE Online. N.B. Text is translated from Dutch.

Nice Online use is limited to subjects having a user account. After logging into the system, ‘standard queries’ are presented to the user. The user can decide to select one of these ‘standard’ queries and choose to either directly view the query’s resulting table or graph or 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 to generate the query. The query model consists of four components: functions, benchmarking/mirror, splitting/intersection, and selection of subpopulation. For each component the user can select from a large list of elements that present either statistical models or data elements in the NICE Registry. Functions allow the user to select the statistics of interest: for example the hospital or ICUs Standard Mortality Rate (SMR). With ‘benchmarking’ the user’s ‘own ICU data are compared to ‘national data’. The splitting/intersection and selection of subpopulation components offer the user the possibility to split the data in gender categories or to create a subpopulation with regard to for example a certain time period, as a few examples. When a user is finished with the query model, he/she can select the ‘graph/table’ button to create the resulting graph or table. The resulting graph or table has to be interpreted by the user himself. Thus background knowledge on statistics is required by users to accurately interpret the resulting data.
At the time of the study a limited number of NICE subjects had requested a user account for NICE Online. In July 2008, NICE Online registered 80 users of 26 of the 79 ICUs participating to NICE at that time. A log file analysis was subsequently performed to gain insight into NICE Online usage patterns. It showed that only 17% of the registered user accounts actually utilized the query functionality on a regular basis; more than 5 times by the quarter of a year. Next, a telephonic information needs analysis provided insight into reasons for the low adoption based on end-user experience with NICE Online. It indicated that users were willing to use the tool but the structured model for query development was considered difficult to use. 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. In addition, planned expansions to the NICE Database required a high level of user-friendliness of NICE Online. This made it necessary to assess its usability and if necessary improve on its functionality. More information on the NICE online system can be found in [21].

Materials and Methods

Subjects

To select a number of representative study subjects, target users were categorized based on a log file analysis of NICE Online usage patterns. Eight experienced users having developed at least 30 individual unique queries in the Tool in the 12 preceding months, and eight new users holding a user account but without use experience with the Tool were randomly selected. For testing of the Physician Data Query Tool formal permission to contact the users was obtained from the NICE Foundation. The selected subjects were then contacted by email with an attached formally signed letter with the NICE foundation approval and the request to participate. All users agreed to participate to the NICE Online evaluation study. In total 16 subjects were contacted.

Tasks

Six usability testing tasks were developed, each composed of two to six subtasks. Input into the development of these tasks was given by two data managers of the NICE Foundation. The NICE data managers are skilled in ICU clinical query development and were able to provide relevant tasks of varying complexity with a gold standard for how to perform and finalize each task supported by NICE Online. The tasks were preceded by a query description, or short clinical question, which could be answered in NICE Online. The formal usability test started with two standard query tasks, randomly given to each subject. Then four tasks with a varying degree of complexity were randomly presented to the subject. These four tasks consisted of two custom query tasks and two tasks consisting of a
clinical question to be translated into a query in NICE Online. In the custom query tasks the subject had to enter query statements in line with the structured query model of the Query tool. In the clinical question tasks a descriptive scenario was given which the subject had to translate to a query in the Query Tool. An example of a custom query task and a clinical question query task is given in Table 1.

Table 1- Examples of the usability testing tasks

<table>
<thead>
<tr>
<th>Examples Main question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Custom query task</strong></td>
</tr>
<tr>
<td>'Please select</td>
</tr>
<tr>
<td>1) the percentage of patient admissions</td>
</tr>
<tr>
<td>2) split by admission type (non surgical, elective surgical, urgent surgical)</td>
</tr>
<tr>
<td>3) for the data of your own hospital,</td>
</tr>
<tr>
<td>4) within a sub selection of the last two years.</td>
</tr>
<tr>
<td><strong>Translating a clinical question task</strong></td>
</tr>
<tr>
<td>'The annual NICE report shows a difference in the mean length of stay for patients in the age category 60 to 70 year in the year 2007 compared to the year 2008. You wish to find out if this is correct by making a graph of these data in NICE Online.'</td>
</tr>
</tbody>
</table>

Usability testing experiment

To characterize the subjects’ user profiles a questionnaire was handed out to each subject before the start of the usability experiment. This questionnaire contained eight questions concerning the subjects’ age, gender, and subjective measurements of their experience levels with computer usage, statistical data processing, and database query demand development (e.g. use of SPSS or Microsoft Access applications). Also the place (at home, at work, both) and manner of use (for research or management purposes) regarding computers, statistical data processing, and development of database queries was inquired. Users with similar profiles were evenly assigned to CTA and RTA in a matched randomized way.

The experiments took place in the actual clinical working area of the subjects. A portable usability laptop with TechSmith’s Morae software (Morae recording tool) allowed recording of 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 testing laptop. Subjects received the tasks in paper form and were given oral instructions on how to carry them out on the laptop. Before testing started, users were asked if they were right or left handed (for mouse configuration).
**CTA condition:** In the CTA condition, subjects were instructed to think aloud while performing the usability testing tasks. In line with the Think Aloud procedure described in [13], it was made clear that the accompanying researcher, the facilitator, would not interfere with the session by giving assistance. The facilitator would remind the subject to keep thinking aloud if the subject would fall silent. Prior to starting the test each subject in the CTA condition received think aloud training tasks to practice verbalizing their thoughts. The training tasks consisted of verbalizing the performing of simple computations.

**RTA condition:** In the RTA condition, the subjects received the same usability testing tasks in paper form. They were instructed to carry out the tasks in support of the Tool, without assistance of the facilitator (silent condition). Directly after subjects were finished, the silent test condition was stopped. Subjects then received the think aloud training task to practice thinking aloud. Subsequently, subjects were asked to verbalize their thoughts retrospectively while viewing the video recording of their actions in the system. In doing so they constructed the retrospective verbal reports. In the RTA condition the Morae recording Tool recorded two files: the user task performance while performing the tasks in the Query tool and the retrospective verbalizations of subjects reviewing their task performance.

**Data analyses**

All audio of the 16 subjects recorded with the Morae software was analyzed and transcribed to verbal protocols. Per subject two usability researchers independently analyzed the time spent in minutes by logging on the time-line of the Morae software. Time spent per subject was calculated as the time between starting a first task in the Tool until the moment of ending a final task in the Tool. The final task was considered completed when the subject mentioned to be finished performing a task (as instructed in the task description). For the RTA condition the time spent on retrospective verbalizations was additionally calculated in the same manner.

Based on the verbal protocols of the CTA and retrospective verbal protocols of the RTA a coding scheme for analysis of usability and task relevant verbal statements was developed. Six randomly chosen verbal protocol transcripts per test condition were first selected and coded. Subsequently all transcripts were coded based on the coding scheme and matched to corresponding video recordings analyzed by two usability evaluators. The same two usability researchers individually categorized the usability problem descriptions into problem types based on the work of Kushniruk et al [22]. Evaluators’ usability coding and categorization consistency was assessed using the Cohen’s kappa statistic. For each method condition a list of detected usability problems agreed upon was then generated. By union of these lists a total set (without redundancy) of usability problems found in the Query Tool was constituted. This overall list is considered to contain all ‘real’ usability problems.
existent in the query tool. Before further analysis could proceed, the average detection rate of usability problems per method was assessed. This measure indicates whether the small sample sizes assessed in the study were sufficient to cover the number of usability problems that potentially could exist in the Query Tool. Next, the frequency (the number of times a usability issue is mentioned) and persistency (the recurrence of usability problems in different tasks) of user experienced usability problems was assessed. Based on these analyses and the potential impact of a problem on task performance a severity rating of 1 (minor) to 4 (high severe) was given to each usability problem, as defined by Nielsen [23]. The severity of usability problems was then rated by two usability researchers and the head software engineer of the NICE Query Tool. To assess the inter rater reliability of three evaluators Fleiss’ kappa was measured [24].

To compare the methods’ performance first their ability to detect usability problems is assessed. Hartsen et al. researched how to balance the validity and reliability of different usability methods in a standardized manner. They assessed specific criteria which UEMs should be able to fulfill in order to be suitable for a specific usability study approach. Some of these criteria are combined in the measure of a methods’ usability problem detection effectiveness, where \( \text{Effectiveness} = \text{thoroughness} \times \text{validity} \) of the method [25]. In this study Effectiveness measurements are analyzed with reference to the time spent per subject in the testing.

Thoroughness refers to the ability of a UEM to find as many usability problems as possible when the user performs tasks with the evaluated system. To measure the thoroughness of CTA or RTA the proportion of the usability problems found by either CTA or RTA to the total of usability problems found by both methods in the Query Tool is assessed. Validity measures whether different UEMs, when compared, find problems that are real problems in system use. Validity of the CTA or RTA method is then measured by the proportion of real usability problems found by a method to the total number of potential usability issues that were found (in first instance) by use of the method. Pearson chi-square test is performed to assess whether a difference in detection effectiveness between methods is significant. In addition to the overall comparison of the methods’ effectiveness their thoroughness was weighted by problem severity levels (1 low to 4 high severe).

Then, the methods’ output in terms of type and severity of problems detected were compared qualitatively. In addition, the verbal protocols were qualitatively analyzed in-depth to explore further reasons for potential differences in verbal protocol output.

**Results**

**General results**
In Table 2 the user characteristics of the 16 subjects are given. Subjects in both method conditions were similar in age, profession and gender. The CTA analysis resulted in 38 potential usability problems of which two were excluded after discussion between the usability evaluators, leading to a total of 36 unique usability problems detected. The RTA found 36 usability issues of which five were not considered to impact the current usability of the Tool but were considered informative to redesign. After discussion between the evaluators a total of 31 unique usability problems were found to be detected by the RTA. After union of the CTA and RTA results in total 43 unique real usability problems were detected in the Query Tool. Inter rater reliability of the usability problem detection consistency was measured $\kappa = 0.93$, $p<0.001$. The 43 usability problems were categorized in seven usability types, $\kappa= 0.83$, $p<0.001$, showing almost perfect agreement among evaluators. Of the 43 usability problems, 20 were classified with a severity 3 or higher. Inter rater reliability of the severity classification of usability problems was $\kappa= 0.69$, $p<0.001$; showing substantial agreement. Table 3 provides an overview of the number of usability problems detected per method.

Based on these first results we additionally assessed whether the small sample sizes were sufficient to cover the potential number of usability problems that exist in the Query Tool at the moment of evaluation. We therefore computed the average detection rate of usability problems per method [26]. The average detection rate of usability problems in the CTA condition was 0.61 (sd 0.07); in the RTA 0.53 (sd 0.06). Usability literature states that for average detection rates higher than 0.45, 8 subjects per method condition provide at least 95% of the existing usability problems in a system [25, 26]. Based on these measured detection rates further analyses of the usability data output was considered reliable.

Table 2. User characteristics of study participants

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>CTA (n=8)</th>
<th>RTA (n=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (sd) in years</td>
<td>38.6 (sd 4.27)</td>
<td>42.3 (sd 6.96)</td>
</tr>
<tr>
<td>Profession</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICU Physician</td>
<td>7 (87.5%)</td>
<td>8 (100%)</td>
</tr>
<tr>
<td>ICU manager</td>
<td>1 (12.5%)</td>
<td>-</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, n(%)</td>
<td>7 (87.5%)</td>
<td>7 (87.5%)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>1 (12.5%)</td>
<td>1 (12.5%)</td>
</tr>
</tbody>
</table>

Table 3. Overview of distinct usability problems detected per method and severity level.
Effectiveness of CTA and RTA

Table 4 shows effectiveness of the two method conditions calculated by their thoroughness and validity. The validity of the RTA was considerably lower than CTA. The five usability issues detected but which did not affect subject task performance mainly caused this difference. In this study the CTA mounted up to 18% higher effectiveness in detecting usability issues compared to the RTA method. Additional Pearson chi-square test showed that the effectiveness of the CTA was indeed significantly higher than the effectiveness of the RTA in revealing unique usability problems ($\chi^2$ 4.54 (1) $P=0.03$).

In both conditions the verbal protocols (concurrent or retrospective) were matched with the video recordings of subject task performance to detect and assess usability problems in NICE Online. To assess the effectiveness of the methods performance the average time spent per subject in both method conditions was calculated. Table 5 gives an overview of the average time spent per subject on task performance and retrospective verbalization. During Task performance the RTA amounted to significant less time effort (22%) for subjects to perform the tasks in NICE Online compared to CTA. The Retrospective verbalization approximately doubled the time subjects spent on the usability testing of NICE Online. This leads to respectively 42% time saving of the CTA pertaining to the 18% higher effectiveness.

With regard to CTA and RTA thoroughness in detecting moderate and severe usability problems a variation in methods performance is also visible (See Figure 2). The difference in thoroughness was most prominent in usability problems of a severity 2 (15% higher effectiveness of CTA) and severity 4 (respectively 0.71 vs. 0.57; a 14% difference in higher effectiveness of CTA). The thoroughness of the RTA weighted per severity level was overall lower than the thoroughness of CTA.
Table 4. Thoroughness, validity and effectiveness per method condition

<table>
<thead>
<tr>
<th>Method condition</th>
<th>Thoroughness</th>
<th>Validity</th>
<th>Effectiveness (Th*Va)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTA (n=8)</td>
<td>(36/43) 0,84</td>
<td>(36/38) 0,95</td>
<td>0,80</td>
</tr>
<tr>
<td>RTA (n=8)</td>
<td>(31/43) 0,72</td>
<td>(31/36) 0,86</td>
<td>0,62</td>
</tr>
</tbody>
</table>

Table 5. Average time spent per subject per method condition

<table>
<thead>
<tr>
<th>Time spent</th>
<th>RTA n=8</th>
<th>CTA n=8</th>
<th>CTA vs RTA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (min, sec)</td>
<td>SD</td>
<td>Mean (Min, sec)</td>
</tr>
<tr>
<td>- Task performance</td>
<td>41.07</td>
<td>7.49</td>
<td>50.16*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Retrospective verbalization*</td>
<td>46.05</td>
<td>5.62</td>
<td>-</td>
</tr>
<tr>
<td>Overall time spend per participant</td>
<td>87.12</td>
<td>-</td>
<td>50.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* During retrospective verbalization a subject was able to stop the recording to verbalize his thoughts. Morae recorded the total time of reviewing the video potentially leading to a longer time on the retrospective verbalization than the actual task performance.

Figure 2. Thoroughness of the CTA and RTA method as a function of the severity
Type of usability problem detected by CTA versus RTA

Figure 3 gives an overview of the distribution in problem types of overlapping and unique usability problems found in the CTA and RTA analyses. Unique usability problems detected per method signify that CTA and RTA usability problem detection scope differed. CTA showed higher sensitivity to usability issues related to the Query Tool’s graphics/symbols, navigation issues, error messages and the layout and organization of the information on the Query Tool’s screens. Of these usability issues detected only by the CTA, three problems were of a severity 4 and three of severity 3 (see Table 6). RTA revealed two unique problems related to understandability of graphics/symbols used in the query tool to generate a query (severity 3). Furthermore, in the RTA condition 4 unique usability problems on terminology and meaning of labels (of which two of severity 4) were revealed. These problems first signified a mismatch between the user (intensivist’s) language and the terminology implemented in the Query Tool. For example, to generate the desired query specific data elements had to be selected by use of search functionality or by browsing through a list of labeled elements. Naming of these labels, such as ‘gender, admission type’ was partly derived from the NICE data-dictionary which contains all data-elements collected by the NICE registry. In both CTA and RTA testing, subjects selected incorrect data elements in several scenarios to generate the queries. During CTA all subjects mentioned the usability problem of incorrect ordering of elements and the organization of the elements on the screen. However, in RTA subjects mentioned that they were unable to deduce the proper use of an element in the query design by means of their label. It appeared that data elements in the Query tool were labeled on differing levels of cognitive complexity; from single elements such as ‘gender’ to more complex elements such as ‘age group per ten years’ (which supports data analysis of patients over periods of ten years). This added to the cognitive complexity of developing a query in the Tool and was remarked upon by seven subjects in the RTA condition. In addition, subjects could select the same type of data elements under multiple components of the query model in the Tool; ‘age group per ten years’ could be selected under ‘splitting/intersection as well as under ‘mirror’. To comprehend how either selection impacts the resulting figure of a query, direct result review becomes a requirement, which was not supported by the Query Tool.
Figure 3. Graphical overview of Type of problems detected per method.

Table 6. Distribution of usability problems per method by problem type and severity

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>CTA/ RTA overlapping* (n=24)</th>
<th>CTA Unique (n=12)</th>
<th>RTA Unique (n=7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CTA</td>
<td>RTA</td>
<td>Severity</td>
<td>Sevety</td>
</tr>
<tr>
<td>Visibility of system status</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Overall ease of use</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Error messages/ help instructions</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Meaning of labels/terminology</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Layout/screen organization</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Graphics/ symbols</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Navigation</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total: 36 31 3 11 8 2 6 3 3 3 2 2

* overlapping = problems detected in both methods of similar nature
In-depth qualitative analysis of CTA and RTA verbal protocols

Furthermore CTA and RTA verbal data showed additional differences in usability problem detection. First of all, though CTA consisted of a higher number of verbal protocols, RTA verbal protocols proved more explanatory towards the more severe usability problems. More precisely, the RTA verbal comments provided additional insight in underlying causes of the occurrence of these severe problems. Subjects in the CTA condition who did not fully comprehend what the term splitting/intersection meant in the query model, showed irritation and commented upon their lack of understandability of the Query Tool terminology. RTA subjects however commented upon the terminology to be ambiguous and provided synonyms to splitting/intersection they perceived would describe it more clearly. Also, subjects who were already familiar with the use of Query Tool prior to usability testing provided in the RTA condition highly detailed information when confronted with a problem they were already familiar with. Usability issues they had experienced before in using the Tool had thus been already thought over by them. During the RTA testing they would specifically address the issue showing how important they thought the problem actually was. In doing so they provided solutions to improve on detected usability issues. Next to solutions the more experienced subjects in the RTA also provided sight on five usability issues which were classified as additional user recommendations for extended system functionalities (utility) of the Tool. For example, one of these recommendations was related to use of the Tool for a deeper scientific analysis of patient cohorts.

The analysis of the number and type of usability problems detected in the CTA and RTA showed that the mental model of subjects, defined by Nielsen as “what the user believes about the system at hand”, did not match the query design model in the Tool [27]. In both CTA and RTA subjects’ expectations and believes on query design in the tool, resulted in severe problems in task performance. Subjects who were not familiar with the Tool as well as experienced subjects encountered the same type of usability issues. This indicated that experienced users had not changed their mental model of query development in the Tool based on their learning experience with the Tool. To extract the mental model of subjects in querying data in the Query Tool their coded articulations on how they intended to solve the usability testing task in the system was further analyzed. The CTA verbal protocols proved more informative for this task. They contained a higher number of verbalizations, provided more insight into usability problems detected and provided more verbalizations on direct task performance compared to the RTA protocols. Therefore the coded intensivist’s verbalizations on query formulation were first modeled for the CTA condition in accordance with their information search behaviour. The RTA coded verbal transcripts were subsequently compared to the resulting query design model to assess differences between subjects information search behaviour in both methods. Results of this analysis showed that overall subjects approached the solving of the Query in the testing tasks in the same convergent manner, for both simple to more complex queries.
The Query model in the Query Tool did not support the query solving approach of subjects. Instead of ‘function’, ‘splitting’, ‘intersection’ and ‘mirror’, subjects started from the basis of what they were supposed to measure, to which data elements they needed to compare or include in the query. Only when they had sight on the data they were interested in, the last phase of the query design model was approached: benchmarking of the data. The CTA results, presented in this study, provided sight on the system requirements and were used as input in redesign of the Query Tool to match its users’ mental model.

Discussion

This study compares the performance of the CTA and RTA method in a usability case study of an ICU Query Tool. The analysis revealed that the effectiveness of the CTA in general and its thoroughness for detecting both moderate and severe usability problems was higher than that of RTA. The methods also differed in their detection scope of usability problem types. The CTA condition in this study in general detected 12 more unique usability issues on graphics/symbols, navigation issues, error messages, layout and organization of information on the Query Tool’s screen, overall ease of use and visibility of system status. The RTA in this study detected more usability problems related to unclear system terminology for users. In addition, RTA offered insight into new system functionalities.

From the field of psychology certain studies provide consistent results to ground this research. Firstly, Kuusela and Paul compared verbal protocol segments of the CTA and RTA approach. They showed that the CTA condition produced a higher number of protocol segments than the RTA condition and elucidated on this difference to be caused by the cognitive nature of the CTA [28]. In contrast to RTA the CTA method appears to evoke subjects to verbalize their short-term memory (STM) task-related thoughts. As verbal protocols in CTA are collected during task performance this leads to a high number of protocol segments. Our study adds to this, by indicating that the difference in cognitive focus of the methods leads to a difference in usability focus. During CTA the verbal protocols reflect STM thoughts on task performance thereby leading to a higher effectiveness of the CTA in detecting usability issues related to the direct user task performance. As a result, usability evaluation studies using only the RTA method may be less able to detect usability problems that make instant requests of the STM of subjects. However, the in-depth analysis of the verbal protocols of the RTA showed a merit of using RTA in formative usability evaluation for the purpose of system redesign. RTA verbal statements consist of a higher rate of user recommendations and are of a more explanatory nature than the CTA. This is in line with prior research of Bowers and Snyder in which they state that RTA verbal protocols consist of more explanations and design statements [16]. However, our research also indicates that subject experience levels could be of influence on this result. In this study, the subjects experienced with NICE Online, and with expertise in clinical data analysis, prior to TA testing were
more informative in providing design solutions and in stating new requirements for additional functionalities of the Tool. Research on expert performance in many domains such as chess and computer programming; have revealed maximal adaptations of experts to domain-specific constraints [29]. Experts input into usability tests might thus differ considerably compared to novice users. However, because of the small number of experienced subjects in the RTA condition a conclusion on this cannot be drawn. Further research into subject computer and system experience on usability testing results of RTA and CTA is therefore required. In this study, even distribution of new and more experienced users in the methods’ conditions prohibited potential bias of subject expertise level on Effectiveness measurements.

Compared to CTA, the retrospective verbalization of RTA may thus evoke a more rational and reflective report on task information processing behaviour as the computer supported task has already been completed and can no longer be changed. These results seem to indicate that though RTA gives sight on unique issues regarding user terminology it might be better equipped for revealing the underlying cause of a usability problem or new user requirements for specific user groups or customized design.

Next to the results of the CTA and RTA comparison, this research also addresses the complexity of clinician querying of large computerized repositories of clinical data variables. This is as far as we know the first usability test on the cognitive complexity of query design by clinicians themselves in a online Query Tool. The foremost problems of the NICE Online tool concerned the comprehensibility of Query model in the Tool and its applied terminology. This mismatch between the Query Tool’s Query model and the mental model of end-users data querying inhibited the system’s convenient use. From the results of the CTA and RTA test a new query model was developed. This new query model could be extracted from the verbal protocols. Additionally, several user requirements were revealed by this study. Among these was the need to provide clinicians the ability of an overall view on all data items in complex data queries combined with direct review of the query result. Following the results of this study the Physician Data Query Tool has been redesigned. A subsequent pre-post study assesses the effect of the Think Aloud testing on the usability of the Query tool’s redesign. In this study the differences in the cognitive complexity of data querying in the old versus the new tool and its effect on the mental workload of data querying by clinicians in the new NICE Online are assessed.

To conclude; though this study indicates that CTA is more effective in usability problem detection the method does not outperform RTA. RTA additionally provides sight on unique usability problems and is of a more explanatory nature than CTA. Based on the results of this study we recommend the use of CTA in formative usability evaluation studies of health information technology. CTA enables the researcher to assess the cognitive processes of subjects during task performance, thereby providing insight into the mental model of users interacting with the healthcare application. The RTA in this study seems to lack the potential to assess the cognitive processes in direct task performance; signified
by less usability output. However, based on the high explanatory value of RTA verbal protocols we recommend RTA in usability studies focusing on user customized (re)design.

Acknowledgements

The authors thank the NICE Foundation for their assistance and support in performing this study. Also all NICE participants who acted as subjects in this study are thanked for their input. We are specifically grateful to the chief software engineer of the NICE registration, Eric van der Zwan, for his suggestions and continuous support in performing this study. The authors also acknowledge Bas de Veer for his contribution as usability evaluator in the performance of the usability testing and subsequent analyses. We thank Joost Dusseljee and Sylvia Brinkman for their input in conducting this study as NICE data managers.

REFERENCES


