Usability Evaluation of a Voluntary Patient Safety Reporting System: Understanding the Difference between Predicted and Observed Time Values by Retrospective Think-Aloud Protocols

Lei Hua[1,2] and Yang Gong[2]

[1] Informatics Institute, University of Missouri, Columbia, MO, USA
[2] School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA
lh5cc@mail.missouri.edu, Yang.Gong@uth.tmc.edu

Abstract. The study evaluated the usability of a voluntary patient safety reporting system using two established methods of cognitive task analysis and retrospective think-aloud protocols. Two usability experts and ten end users were employed in two separated experiments, and predicted and observed task execution times were obtained for comparison purpose. According to the results, mental operations contributed to the major effort in reporting. The significant time differences were identified that pointed out the difficulty in human cognition as users interacted with the system. At last, the data collected by retrospective think-aloud technique, e.g. the response consistency on structured questions and the user's attitudes, revealed the frequent usability problems impeding completion of a quality report.

Keywords: patient safety, voluntary reporting, cognitive task analysis, retrospective think-aloud.

1 Introduction

The Institute of Medicine called for nationwide reporting systems to collect medical incidents for patient safety improvement in 1999, the year when "To Err is Human" report was released [1]. It is believed that the reporting systems would be a data source to learn from the lessons, if safety events were collected in a properly structured format for the detection of case patterns, discovery of underlying factors, and generation of solutions. Since 2008, 26 States had implemented hospital medical error and incident reporting systems [2]. However, there are gaps between the status quo and the potential of the reporting systems, because of the challenges in user engagement [3] and data quality[4, 5]. As a critical contributing factor, usability has received little attention in dealing with the challenges.

In this study, we employed two usability methods of cognitive task analysis and retrospective think-aloud protocols to evaluate a patient safety reporting system. The difference between predicted and observed time from two experiments drew attention to the sites where the user's performance was significantly affected. The analysis of task responses and think-aloud protocols helped identify usability problems and their underlying factors at the sites.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2013, LNCS 8005, pp. 94–100, 2013.

2 Methods

2.1 The Study System

The development of the tested system was based on the navigational structures of an implemented reporting system in a local health organization [6]. It implemented the Common Formats (CFs) for collecting case details. Developed by the Agency for Healthcare Research and Quality (AHRQ), the CFs aim to diminish the disparity of categorizing and describing patient safety events among the existing patient safety organizations and reporting systems. For each event category, CFs offer a standardized list of multiple-choice questions (MCQs) to promote case reporting.

2.2 Cognitive Task Analysis, GOMS and KLM

Cognitive task analysis (CTA) is a widely used usability evaluation method to describe the tactics and knowledge that underlay task performance. The method employs usability experts and GOMS (Goals, Operators, Methods, and Selection rules) model to examine the user's physical and cognitive steps and barriers in the task execution. For the measures of execution time and mental-physical ratio, Keystroke Level Model (KLM) was used to estimate mental and physical operations in seconds. It refers to seven operators with estimated execution time on each.

- K – Keystroke : 0.28 Sec
- T (n) - Type a single chuck of n characters in a sequence on a keyboard : n*K Sec
- P - Point with mouse to a target on the display : 1.1 Sec
- B - Press or release mouse button : 0.1 Sec
- BB - Click mouse button : 0.2 Sec
- H - Home hands to keyboard or mouse : 0.4 Sec
- M - Mental act of routine thinking or perception : 1.2 Sec

Differing from the peers under the GOMS family [7], GOMS-KLM considers the individual operations in a linear sequence and sum them up for predicted execution time as shown in Table 1. In the study, the predicted time served as a baseline of reporter's performance, to pinpoint the observed data that significantly varied from the prediction.

Table 1. "Entering occurrence time of an event" subtask using GOMS with KLM technique

	GOMS		KLM	
Step #	Step description	Distributed cognition Physical/Mental operator	Operators	Time (s)
Step 1	Locate the field for date entry	Mental	M	1.2
Step 2	Point the mouse to the field	Physical	P	1.1
Step 3	Click to put the cursor into the field	Physical	B	0.1
Step 4	Verify the date field that obtains the focus	Mental	M	1.2
Step 5	Hand keyboard	Physical	H	0.4
Step 6	Retrieve the date	Mental	M	1.2
Step 7	Interpret the date value into required format	Mental	M	1.2
Step 8	Type the formatted date	Physical	T(10)	2.8
Step 9	Verify the date and its format are correct	Mental	M	1.2
Step 10	Home hand to mouse	Physical	H	0.4
			Total	10.8

2.3 Retrospective Think-Aloud Protocols

We applied the retrospective think-aloud (RTA) to measuring user's performance in aspects of execution time, data quality and user's attitudes. The method asked participants to verbalize their thoughts after the reporting session activity, instead of during the session. The method avoids obtrusive task disturbances that were usually introduced by concurrent think-aloud technique to the performance.

2.4 Participants

Two usability experts and ten end users were recruited for the CTA and RTA experiments separately. In RTA, the invitation letter and screening form were emailed to the School of Nursing and the School of Medicine at the University of Missouri for qualified participants. The qualified respondents were those who had reported patient falls at least once and were interested in online patient safety reporting systems. The first ten available candidates became the testing participants. Every study participant was required to sign on an informed consent form, according to the approval of the Institutional Review Board in the university.

2.5 Task Scenarios and Testing Steps

The task was to report three patient fall events in the system. Three fall cases in a written format were selected from a library of 346 fall reports. The cases were reviewed by domain experts to ensure quality and readability. Fall event cases were chosen for the test because the fall reporting form in the CFs is simple and structurally representative, and falls are typical in hospitals at all levels. An example of a fall event scenario selected from the library is shown in the following excerpt:

> *… the patient indicated need to be toileted. He stood with a walker and walked to the bathroom. He noted less steady than yesterday, dragging right leg. He turned while in the bathroom toward the sink …*

Table 2. Time performance and material accessibility by subtask

Subtask	*Task name*	*Time (s)*	*Access to written materials*
#1	Answer initial questions	18.3	Yes
#2	Rate a harm score	28.1	No
#3	Enter patient related info	100.8	Yes
#4	Answer structured MCQs	102.2	No
#5	Document further comments	34.5	No
	Total	283.9	

In both experiments to fulfill a reporting task, the participants needed to complete five subtasks sequentially as shown in Table 2. In practice, the reporters at the work site documented case-specific information upon memory. Thus, in a simulated setting as it was in the RTA test, the participants were not allowed to review the written materials for completing case-specific subtasks #2, #4 and #5, once did the reporting start.

In CTA, GOMS was performed on the set of five tasks to identify common task steps. Two evaluators (LH and RG) independently conducted GOMS on each of the five tasks. Inter-rater reliability was calculated to determine the extent to which two evaluators agree with each other on dividing task steps and assigning physical/mental operators.

In RTA, the ten participants were assigned separate time sessions for the test. They were trained by a video demonstrating how to manipulate the system for completing a report. Each session was audio and video recorded using Camtasia Studio® 7. Ten participant's task performance and verbalization were collected for data analysis.

2.6 Processing of the Data

For the purpose of comparison, we focused on time performance in the two experiments. Predicted and observed execution times were collected from the two experiments and then contrasted by tasks and subtasks. The time performance in CTA consists of two parts. The sum of six physical operators' time on the task represents user's physical execution time, and the amount of mental operators involved determines user's mental execution time. These predicted time values were served as a benchmark to contrast with counterparts observed through RTA, in which the observed execution time was split into two parts based on the collection of physical operators and execution times by the session review. The difference between the predicted and observed time served as an indicator of the system usability problems that users encountered in RTA.

Since the usability problems might have negative effects on the quality of reports and user's attitudes, related data were collected for the evaluation. The response consistency on structured questions was calculated by generalized Kappa[8], to account for the kind of easiness that users were able to reach a consensus. A low consistency inferred the existence of usability issues on the question. In addition, the participants' think-aloud verbalizations were transcribed and coded by a scheme developed by Zhang et al[9]. The coding scheme comprised 14 usability heuristics assisted in classifying usability problems that influenced participant's performance in the test. Any disagreement in classification was resolved in discussions among research team members until a full agreement was reached.

3 Results

In CTA, the mean counts of task steps was 225 that consisted of 93 physical and 132 mental operations. In total, a report took 266.6 seconds averagely for a report and 108.2 seconds and 158.4 seconds respectively. The ratio of mental/physical operators was 58.67% as shown in Table 3.

In RTA, the mean of reporting completion times is 277.9 seconds. 102 physical operators were involved for each report and accounted for 96.5 seconds. The difference between total and physical times of 181.4 Sec is construed as the mean of actual mental times on a report. All above results are listed in Table 4.

Table 3. Time performance and material accessibility by subtask

Task #	Task name	Total steps	Operators		% Mental	Est. time (s)		Time in total (s)	Kappa
			Mental	Physical		Mental	Physical		
1	Answer initial questions	22	11	11	50.00%	13.2	6.9	20.1	0.937
2	Rate a harm score	14	10	4	71.43%	12.0	2.4	14.4	0.606
3	Enter patient information	103	56	47	54.37%	67.2	32.4	99.6	0.888
4	Answer structured MCQs	72	49	23	68.06%	58.8	34.2	93.0	0.802
5	Document further comments	14	6	8	42.86%	7.2	32.3	39.5	0.651
	Total	225	132	93	58.67%	158.4	108.2	266.6	

Table 4. User Testing with KLM and think-aloud technique

Task #	Task name	Total Time(s)	Observed Physical		Mental time (s)		Diff.
			Operators	Time(s)	Obs.	Pred. in Table 3	
1	Answer initial questions	18.3	11	11	11.4	13.2	-13.64%
2	Rate a harm score	22.1	10	4	19.7	12	64.17%
3	Enter patient information	100.8	56	47	65.5	67.2	-2.53%
4	Answer structured MCQs	102.2	49	23	76.9	58.8	30.85%
5	Document further comments	34.5	6	8	7.9	7.2	9.72%
	Total	277.9	132	93	181.4	158.4	14.52%

Considering CTA results as a benchmark, the majority of observed execution times from RTA were within the error limit of +-21% suggested by GOMS-KLM[10]. Task #2 and #4 were exceptional as shown in Table 4. We thus looked into them at the single question level as subtasks as shown in Table 5. Half of the subtasks (6 out of 12) were beyond the limit that took either much less or more time than the prediction. The agreement of the choice selection on each of the subtasks was calculated and attached except for subtask #4.9 that allows checking multiple choices for the answer.

Table 5. Comparison of estimated and actual mental time on subtasks of task #2 and #4 (multi-choice questions), with agreement rate of 10 subjects' choice selection

Subtask #	Subtask name	# of choices	Mental time (s)			Generalized Kappa
			Obs.	Pred.	Diff.	
2.1	Rate a harm score	6	19.7	12	64.17%	0.385
4.1	Q(1) Assisted fall or not	3	3.57	3.6	-0.93%	0.748
4.2	Q(2) Observed fall or not	3	1.97	3.6	-45.37%	0.867
4.3	Q(3) Observed by who	2	2.68	3.6	-25.51%	0.719
4.4	Q(4) Patient Injured or not	3	3.23	3.6	-10.19%	0.933
4.5	Q(5)* Type of injury	5	9.00	4.8	87.58%	1.000
4.6	Q(6)* Prior doing ahead of falling	11	12.45	4.8	159.31%	0.304
4.7	Q(7) Fall risk assessed or not	3	7.41	3.6	105.74%	0.363
4.8	Q(8) Patient at risk or not	3	3.95	3.6	9.72%	0.833
4.9	Q(9)*~ preventive protocols	16	24.76	20.4	21.37%	N/A
4.10	Q(10) Med increased risk or not	3	4.06	3.6	12.78%	0.630
4.11	Q(11) Med contributed to fall or not	3	3.87	3.6	7.41%	0.696
			76.9	58.8	30.85%	

In the think-aloud protocols, fifty-seven comments were coded into nine categories of usability problems reflecting user attitudes. Some comments that referred to multiple categories were categorized into the best fit. The most frequently identified problem was the language problem – 15 comments (26.3%) and every subject had at least one comment on CFs questions. The common issues were match (22.8%), memory (15.8%), visibility (12.3%) and feedback (8.8%). Most of the coded problems in the top five categories were commenting on cognitive difficulties that subjects encountered in the task completion process.

4 Discussion

In two experiments using different usability techniques, three types of data were collected with respect to the reporting time, consistency and user's attitudes. Supposing the time variables from CTA as a benchmark, the comparison identified several significant differences between the prediction and observation. The data regarding the response consistency and user's attitudes from RTA accounted for the underlying factors that might lead to the differences.

Overall, the predicted and observed execution times for a report completion were very close. All time differences regarding physical and mental operations were under the error limit regulated in KLM for time prediction. It indicated that the unknown disturbances, if the RTA had, did not influence the execution times in the observation significantly in comparison of the predicted values.

To complete a report, 93 physical operators were predicted comparing 102 operators in the observation. Not in an ideal circumstance as the testers in CTA that had no hassles on unpredictable redo and typo, the ten participants in RTA might need extra keystrokes or mouse clicks in the real context.

On the other hand, the differences of mental execution times between the two experiments exceeded the error limit on task #2, #4 and some corresponding subtasks as shown in Table 5. For example, the percent variations were 159.31% and 105.74% on the subtask #4.6 and #4.7. Meanwhile, the low responding consistency (considering 0.600 as a dividing threshold [11]) might occur accordingly. It indicated in a few of subtasks reporting case details, extra mental operators and user errors were introduced for unpredicted problems in human cognition. According to the coded comments, usability problems of language, information mismatch, visibility and feedback dominated the cognitive issues that burdened the participants and lowered the participant's performance.

5 Limitation

The findings were based on a specific domain and obtrusive study techniques that might limit the generalizability of identified problems and user's performance in a natural context. To make a comparison of mental execution times between two experiments, we subtracted estimated time of physical operators from the total to obtain the mental time values based on an arguable assumption. It assumed the estimated

execution times of physical operators by GOMS-KLM were accurate and physical and mental operations in RTA could be treated in a linear sequence of execution.

6 Conclusion

The study showed that mental operation accounted for the majority of effort in a report using the system. The mental effort could be affected by usability problems as a reporter interacted with the system interface that slowed the process and undermined the quality of reporting. Cognitive task analysis and think-aloud user testing was helpful to identify these problems and pave the way towards the system usability enhancement.

References

1. Rosenthal, J., Takach, M.: 2007 guide to state adverse event reporting systems. National Academy for State Health Policy (2007)
2. Levinson, D.R.: Adverse events in hospitals: state reporting systems. US Department of Health and Human Services, Office of the Inspector General Washington, DC (2008)
3. Kim, J., Bates, D.W.: Results of a survey on medical error reporting systems in Korean hospitals. Int. J. Med. Inform. 75(2), 148–155 (2006)
4. Gong, Y.: Data consistency in a voluntary medical incident reporting system. J. Med. Syst. 35(4), 609–615 (2009)
5. Gong, Y.: Terminology in a voluntary medical incident reporting system: a human-centered perspective. In: Proceedings of the 1st ACM International Health Informatics Symposium, pp. 2–7. ACM, Arlington (2010)
6. Kivlahan, C., et al.: Developing a comprehensive electronic adverse event reporting system in an academic health center. Jt. Comm. J. Qual. Improv. 28(11), 583–594 (2002)
7. John, B.E., Kieras, D.E.: The GOMS family of user interface analysis techniques: comparison and contrast. ACM Trans. Comput.-Hum. Interact. 3(4), 320–351 (1996)
8. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
9. Zhang, J., et al.: Using usability heuristics to evaluate patient safety of medical devices. J. Biomed. Inform. 36(1-2), 23–30 (2003)
10. Card, S.K., Moran, T.P., Newell, A.: The psychology of human-computer interaction (1983), http://books.google.com/books?id=JeFQAAAAMAAJ
11. Devore, J.L.: Probability and statistics for engineering and the sciences. Brooks/Cole Pub. Co., Monterey (1982)