Heuristic Evaluation of Mission-Critical Software Using a Large Team

Tim Buxton, Alvin Tarrell, and Ann Fruhling

University of Nebraska Omaha, Peter Kiewit Institute - Office 174C
6001 Dodge Street Omaha NE 68182
{ebuxton,atarrell,afruhling}@unomaha.edu

Abstract. Heuristic evaluation is a common technique for assessing usability, but is most often conducted using a team of 3-5 individuals. Our project involved a team of 16 stakeholders assessing usability of a mission-critical decision support system for the US military. Data collected from so many evaluators could easily become overwhelming, so we devised a method to first filter evaluations based on agreement between evaluators, and then further prioritize findings based on their individual Frequency, Impact, and Severity scores. We termed our methodology the 'Integrated Stakeholder Usability Evaluation Process,' and believe it will be useful for other researchers conducting similar research involving heuristic evaluations with large groups.

Keywords: Usability Evaluation Methods, Heuristic Evaluation, Decision Support.

1 Introduction

This research involved evaluation of a recently redesigned user interface for a mission-critical decision support system for the U.S. military. This system is very complex, and is built to handle large volumes of data and account for constantly changing operational conditions. As such, the user interface for this system must provide immediate situational awareness, visual cues to high-priority events, and decision support functionality to a wide variety of users, often under conditions of extreme time pressure and stress. Additionally, the user interface must be as intuitive as possible, assist in error prevention, and require as little training as possible. In short, the usability requirements for this mission-essential system were much more critical than for many applications, and a well-designed user interface is a key contributor to meeting those usability requirements.

Usability is a key to making systems easy to learn and easy to use [1]. Usability includes the consistency and ease with which the user can manipulate and navigate, clarity of interaction, ease of reading, arrangement of information, speed, and layout. Usability improves the design of user interfaces by evaluating the organization, presentation, and interactivity of the interface [2]. Prior research overwhelmingly suggests that usability is associated with many positive outcomes, such as a reduction in the

J.A. Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2009, LNCS 5613, pp. 673–682, 2009.

number of errors, enhanced accuracy, a more positive attitude on the part of the user toward the target system, and increased usage of the system by the user [3].

Cummings and Guerlain [4] describe concerns for cognitive load and time pressures in use of mission-essential and time-critical software, a telemedicine system. Usability problems which increase cognitive effort or increase time to complete tasks can lead to failures of system effectiveness. As a decision support tool with access to large volumes of current information, the system must also function well in Information Retrieval (IR) as described by Xie [5], but as she notes, strategies for IR may change when there is extreme urgency. The system being evaluated shared these characteristics, driving a change from standard usability testing methodologies.

This unique environment necessitated going beyond standard usability testing. We wanted to utilize a User Interface Evaluation (UIE) tool that would contribute to finding usability problems in our situation, where cognitive load and urgency or operational tempo are such major issues Our research method, given below, draws on experiences of others with similar usability needs, and led to some significant results that we describe.

The contribution of this study is the repeatable process that we developed to effectively capture feedback from various key stakeholders on the quality of the system we studied. We adapted the *heuristic evaluation* method and broadened the scope to include UI guidelines established by the CCDS development team. We felt it was the best method, and had in fact been found by Cummings and Guerlain to predict usage problems even for applications used under time pressure [4]. The process is more fully explained in the following sections.

2 Background

Developers create applications providing the required information and functionality, but developers often don't know how to present it in the most intuitive (to domain experts), usable way. Domain experts who use programs may be frustrated by what is in fact a usability problem, but don't realize how easily it could be remedied if identified [6]. Human-Computer Interaction (HCI) professionals have developed systematic ways to bridge this disconnect and improve application usability. When application usability is improved, substantial benefits result to the domain expert users [6]. These come from improved learnability, visibility, user control, error prevention and recovery, and speed of task completion (efficiency).

The problem we set out to solve was: "How can domain experts, HCI professionals, and IT personnel collaborate to find usability problems and make an existing application more usable by applying collective knowledge that no one person possesses?" A second question, given that the evaluation team now consisted of 16 diverse stakeholders, was: "How can we accommodate the large amounts of data generated, how can we define agreement between reviewers, and how can we rank order our findings?"

Our Integrated Stakeholder Usability Evaluation Process was designed to help provide a solution to these questions.

The system studied will be referred to as the Command and Control Decision Support (CCDS) System. CCDS is a strategic information system intended to enhance

command and control of critical military units, providing national-level government leaders and military commanders the capability to monitor the status of critical forces, to make and implement decisions, and to be aware of potential threats.

CCDS command and control capability relies on information made available by several reporting systems. CCDS-processed information forms the basis for decision-making and for implementing decisions about command and control functions. The primary CCDS users are the decision advisors; i.e. the members of the military operations team providing advice to senior decision makers. They obtain CCDS data and information, integrate those with relevant material from other sources, and present the resultant knowledge and recommendations to the key decision makers.

3 Research Method

We selected the heuristic evaluation method as being most suitable because it can be used for existing systems, takes little user time, and is relatively quick and inexpensive. Tang et al. [7] have also demonstrated that heuristic evaluation is able to point to usability problems in a system that required task performance in a timely fashion, leading further credence to its use in this situation.

Jeffries et al. [8] describe heuristic evaluation as having a user interface expert or a group of experts with knowledge of good user interface design principles study an interface and, based on experience and training, identifying potential areas of difficulty. The evaluators are generally experts in usability, although it is often desirable to use individuals who are both usability and domain experts [9]. They study the interface in depth and look for properties that they know, from experience, will lead them to problems. The idea is that while no one individual assessor will find all the violations of the heuristics, several expert evaluators working independently may be very effective. The end result of the evaluation is a list of problems or conflicts with the associated heuristics referenced [10]. When all the evaluators have finished their evaluation they will aggregate their list of problems [11]. The heuristic evaluation method is often selected because it is a cost effective method for an organization that does not have the facilities, time, and expertise necessary to do exhaustive usability engineering.

4 Heuristic Evaluation Process

The Research Team followed a new framework for interface usability evaluation which consisted of six steps detailed in Figure 1 below, starting with the selection of the usability evaluation method and ending with a set of prioritized improvement recommendations. The support tools selected for the usability evaluation were simple tools that were used to manage the process and collect data, in effect extracting the consensus of the group. This allowed for easy and quick iterative refinement throughout the process. We discuss the six steps next.

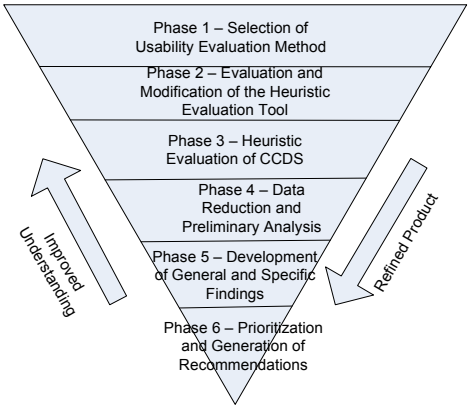

Fig. 1. Heuristic Evaluation Framework

Phase 1 – Selection of Usability Evaluation Method
We were asked to concentrate on the usability of the user interface, and the heuristic evaluation technique works well for that purpose. This approach also did not infringe significantly on end-users' time commitments. However, it is still a thorough, comprehensive process, and it is cost effective and could be completed in a timely manner.

Phase 2 – Evaluation and Modification of the Heuristic Evaluation Tool
Nielsen's Ten Usability Heuristics [12], derived by Jakob Nielsen from a factor analysis of 249 usability problems, were used as the basis of the study. We also included an additional three heuristics identified by Denise Pyrite, Xerox Corporation [13], adding 43 evaluation items. Finally, we also incorporated 37 CCDS-specific requirements contained in the Graphical User Interface guide written specifically for CCDS (CCDS Java User Interface Standards). The resulting thirteen heuristics contained a total of 329 individual evaluation items from a combination of these three sources.

Modifications to the heuristics themselves were also made, including changing the 'Yes/No' structure of the questions to a more appropriate 7-point Likert scale to afford the evaluators some flexibility in determining not only whether system had met the requirements of a particular heuristic, but also how well the system met that requirement. This also provided the opportunity to do more quantitative analysis of the resulting data.

Phase 3 – Heuristic Evaluation of CCDS
This phase began by discussing the meaning of the heuristic and how it might apply to CCDS, then making a determination of whether or not the heuristic should be adopted and whether or not it should be specifically included in the CCDS Java User Interface Standards. If the checklist question was deemed applicable, each evaluator entered a rating on how well CCDS enforced the checklist question. Another column was added for comments by each evaluator that indicated where the system violated the question

and also where the system could be improved. The evaluation process was fully explained to the evaluation team at the start of the research, with a short reminder session provided at each meeting. We also created a comprehensive numbering system for items on the user interface menu tree, e.g. 1.1.2, that included all possible menu selections. This provided a "common format for documentation" as recommended by Koutsabasis, et al. [14].

Because of the mission-critical nature of the software, we wanted as many reviewers as possible. The team consisted of 16 individuals, although not all were able to complete all 13 heuristics. We were able to get 10 reviewers for many heuristics, which should be expected to detect 85% of usability problems, as opposed to 60% for 3 reviewers, according to Nielsen and Landauer [15].

The CCDS Usability Evaluation Team evaluated the system primarily in group sessions. This allowed us to discuss examples of success or failure, and in some cases decide that a heuristic did not apply. If the heuristic was adopted and applied, we navigated through the system evaluating how well the system complied. Each evaluator entered his or her rating independently and also added comments to explain the rating, especially if the system did not comply with the heuristic.

The CCDS application was projected for all to view during most of these discussion sessions, and each person also had the CCDS application available at his or her desktop. The team also captured other enhancement ideas that were generated during the analysis and discussion. Sixteen individuals completed some portion of the evaluation, with six individuals completing all 13 heuristics and the remainder completing the heuristics to varying degrees. A total of 100 person-heuristics were ultimately completed, or an average of 7.7 evaluations per heuristic.

The process we followed is shown in diagrammatic form in Figure 2.

Phase 4 – Data Reduction and Preliminary Analysis

After the evaluations were completed, the data were migrated from the Excel spreadsheets into a Microsoft Access database for query-based analysis and ease of report generation. Forty-six of the original 329 checklist items were rejected during the evaluation sessions as not being applicable to CCDS, so a final set of 283 items were analyzed. Preliminary evaluations for inconsistencies (standard deviations) were first conducted, and then items were first prioritized based on the mean scores. This research produced 2,612 data points (individual heuristic question items rated) and 944 separate comments from the evaluators.

Average heuristic ratings and standard deviations for each of the 283 items were calculated and various reports were generated for analysis. Some of the reports utilized were:

Adopted Heuristics – Complete listing of the heuristics adopted by the research team, sorted by category and item number taken from the original Xerox tool. Additional CCDS User Interface Guidelines were added in under the appropriate heuristic category, on the same line as the Xerox heuristic if it coincided, and on a new line if it did not map to a Xerox heuristic.

Average Ratings by Heuristic – The Adopted Heuristics report, with the addition of average CCDS ratings as assigned by the research team

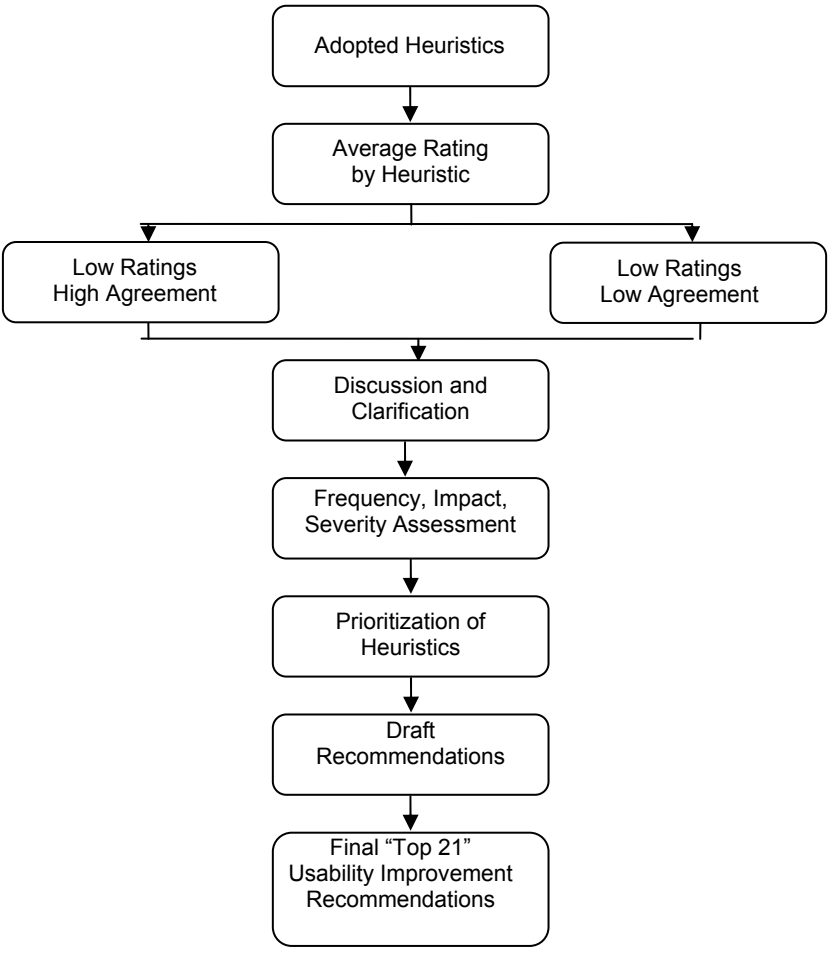

Fig. 2. Steps in the Integrated Stakeholder Usability Evaluation Process (ISUEP)

Low Ratings, High Agreement – A subset of the Average Ratings by Heuristic report, filtered to show only those items that received an average score of less than five on the seven point rating scale with sufficiently low standard deviation as to indicate agreement between the raters.

Disagreement Between Raters – A listing of adopted design heuristics, regardless of average rating, where the standard deviation between the ratings was greater than two. This report was used as a review and revision tool by the research team to identify areas where there might be differences in interpretation or experience with the CCDS system.

Phase 5 – Development of General and Specific Findings
Using the "Low Ratings, High Agreement" report as a starting point, the team identified 70 items that warranted further examination. These were, in general, items that scored an average score of '5' or less on the 7-point Likert rating scale.

A more quantitative analysis was then undertaken to help identify priority items within the group of 70. Each of these items was scored based on how often it occurred (Frequency), how many users were affected (Impact), and the difficulty of recovering from or overcoming the problem (Severity) as described by Nielsen [16]. This discussion compensated for the tendency to produce false positives by HE noted by Hornbaek and Frokjaer [17] by dismissing deviations which had only minimal impact. Relative rankings of the 70 items of interest were developed based on these scores (with a double weighting of the Severity factor), and a prioritized list of potential action items began to take shape. The team was reduced at this point to the three key researchers and the development team representative; we felt this was necessary to allow the group to better focus on the rankings without facing debate among 16 separate evaluators.

Recommendations for corrections and improvements were then developed, often using the original heuristic as a style guide, as well as incorporating CCDS-specific language and ideas. The team also reexamined those heuristics which were just "below the line" for initial consideration, either by average score or by standard deviation.

Phase 6 – Prioritization and Generation of Recommendations
Based on the process described above, the evaluation team identified a prioritized list of twenty-one recommendations to present to the CCDS managers and development team. That resulted in the generation of a summary report:

Final Usability Improvement Recommendations – A "Top-21" list of recommendations presented in order of importance based on of the highest average perception of positive impact to CCDS system usability.

5 Results

This research produced 2,612 data points or individual heuristic ratings and 944 comments. Ranking the least followed heuristics by percentage of items in the Top 70 by Heuristic Category, we found the most problematic heuristic was Consistency and Standards, followed by Flexibility, Error Prevention, Flexibility, and Error Recognition and Recovery.

We observed a number of surprising benefits in our project that can be expected in any heuristic evaluation. Usability professionals initially constitute novice users, and so can be useful in finding misleading parts of the UI. Landauer [18] points out that testing with novices is important to good *user centered design*, and in fact we confirmed this by several catches of usability problems by novices that had escaped experienced users. For example, new users noticed that the function of icons was not clear to new users, and they needed to indicate their function more intuitively and make them more distinct. Some boxes which were shaded as if not selectable were in fact selectable – new users were good at noticing this sort of usability problem.

We also observed that our sessions led to general comments and ideas which were captured with a list of Extra Comments not tied to specific heuristics details. Evaluation sessions often led to serendipitously finding functional errors in the new revisions of the software as well, leading to generation of software deficiency reports. For example, when all fields were left blank on a report design form, the database crashed. So it is reasonable to point out these bonuses to the organization as costs of the method are assessed.

Formal results provided to CCDS included a prioritized list of the 'Top-21' usability items that the evaluation team thought should be addressed. The report also included usability scores and comments for all 283 evaluation items, allowing the development team to use this information as they saw fit.

The developer in our team showed increased awareness of usability best practices as a result of his participation. We would expect that this leads to better usability being built in, as was also observed by Tang, et al. [7].

Most importantly, however, we developed and documented a repeatable process that can be used for usability evaluations using large numbers of evaluators. Our methodology allowed us to quickly assess inconsistencies in evaluations among the different inspectors, look for evaluations with simultaneous low scores and high agreement among the evaluators, and then quickly prioritize those findings based on frequency, impact, and severity.

6 Implications and Conclusion

Our first question of how to apply collective knowledge that on one person possesses to improve usability, we solved in two ways. One was the rating data collection system, including the comments. The small group reading the comments and ratings had access to a lot of pooled collective knowledge to use in usability improvement. An unexpected result of doing the evaluations in group sessions was that discussion always led to participants hearing about problems or perspectives that were "new" to them, so the collective knowledge was increased. We recommend the group sessions for just that reason, over evaluation by individuals asynchronously, if that is possible.

Heuristic evaluators can produce divergent and varied results. Our contribution was to create analytic reports to extract the consensus of the large group that we had engaged to evaluate usability. These reports were then very productive in guiding the discussion of a small team tasked with finding the low-hanging fruit in the usability improvement process. Our assignment of a menu tree numbering system to the application was found to be vital in clarifying discussion of which exact parts of the large application showed specific usability problems, and in communicating to the development team exactly what needed to be done to improve usability.

Heuristic evaluation is just one method for user interface usability evaluation. We developed the Integrated Stakeholder Usability Evaluation Process as a means of dealing with a large group of diverse evaluators; this process is designed to quickly identify items with strong user agreement and with high priority for correction. We recommend that additional usability evaluations - cognitive walkthroughs, task analysis, and user observations – also be considered. These future evaluations could be conducted as stand-alone tests, or could possibly be undertaken during normal CDSS

operations. This would require a minimal commitment of user time, and allow "real life" evaluations, potentially leading to valuable improvements.

Heuristic evaluation is just one strategy for user interface usability evaluation. We recommend that additional usability evaluations be considered such as cognitive walkthroughs, tasks analysis, and user observations. Task completion times, user errors and other evaluations and suggestions could be potentially built in to ongoing routine exercises and after-action evaluations of performance.

Applications are complex, and we found the menu tree numbering system was important to add to clarify both our discussions in evaluating the software, and later for software developers to find and fix the exact problem, without having to relocate it.

Future directions for research might include evaluating the measures of usability problem impact. Are frequency, impact and severity the best scales to use, and how is it best to weight them against each other? Are there other measures of impact for heuristic evaluation that would be of more value in time-critical command and control applications, perhaps prioritizing serious errors, and how likely they are and how difficult to recover from quickly? Also, in large applications, coverage is difficult to assess – even a larger number of evaluators have a limited time, and some areas of the application may still not get tested. There may be ways to improve coverage that could be found in future work.

We applied all of the heuristics for thoroughness. However, we would expect that in a time-critical decision support system such as this, certain heuristics would be found to be more likely to find usability problems that show up during time-critical operation, and time could be saved by applying only those found to be most critical. These would likely be heuristics that add to intuitive "information scent" as described by Spool et al. [19] and "information foraging" as described by Pirolli [20], as being crucial to quickly giving users intuitive guides to where to go to accomplish their intended tasks. Nielsen's *Match Between Systems and the Real World* heuristic might be expected to be key here. Improvements in intuitive usability, in addition to providing more efficient and accurate task completion, might even lead to less training being necessary. Feedback from students during training could be used to improve the intuitiveness of the design, and would be an avenue of further research.

Acknowledgements. We would like to thank the participants in our heuristic evaluation sessions for CCDS at USSTRATCOM. Without their commitment the analysis would not have been as comprehensive or complete.

References

1. Nielsen, J., Mack, R.: Usability inspection methods. John Wiley & Sons, San Francisco (1994)
2. Shneiderman, B.: Designing the User Interface Strategies for Effective Human-Computer Interaction. Addison-Wesley, Reading (1998)
3. Lecerof, A., Paterno, F.: Automatic support for usability evaluation. IEE Transactions on Software Engineering 24, 863–888 (1998)
4. Cummings, M.L., Guerlain, S.: An Interactive Decision Support Tool for Real-time In-flight Replanning of Autonomous Vehicles. In: AIAA 3rd "Unmanned Unlimited" Technical Conference, Workshop and Exhibit, September 20-23 (2004)

5. Xie, H.: Interactive Information Retrieval in Digital Environments. IGI Publishing, Hershey (2008)
6. Nielsen, J.: Usability engineering at a discount. In: Smith, G., Salvendy, M.J. (eds.) Designing and Using Human-Computer Interfaces and Knowledge Based Systems, Elsevier Science Publishers, Amsterdam (1989)
7. Tang, Z., Johnson, T.R., Tindall, R.D., Zhang, J.: Applying Heuristic Evaluation to Improve the Usability of a Telemedicine System.1. Telemedicine and e-Health 12, 24–34 (2006)
8. Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.M.: User interface evaluation in the real world: a comparison of four techniques. Communications of the ACM, 119–124 (March 1991)
9. Kantner, L., Rosenbaum, S.: Usability Studies of WWW Sites: Heuristic Evaluation vs. Laboratory Testing. In: Proceedings SIGDOC, pp. 153–160 (1997)
10. Leventhal, L., Barnes, J.: Usability Engineering, Process, Products, and Examples, p. 216 (2008)
11. Preece, J., Rogers, Y., Benyon, D., Holland, S., Carey, T.: Human-Computer Interaction, p. 676. Addison-Wesley, Reading (1994)
12. Nielsen, J., Molich, R.: Heuristic Evaluation of User Interfaces. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 249–256 (1990)
13. Weiss, E.: Making Computers-People Literate. Jossey Bass (1993)
14. Koutsabasis, P., Spyrou, T., Darzentas, J.: Evaluating Usability Evaluation Methods: Criteria, Method and a Case Study. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4550, pp. 569–578. Springer, Heidelberg (2007)
15. Nielsen, J., Landauer, T.K.: A mathematical model of the finding of usability problems.Amsterdam. In: The Netherlands Proceedings ACM/IFIP INTERCHI 1993 Conference (1993)
16. Nielsen, J.: Severity Ratings. Jakob Nielsen's Website, http://www.useit.com/papers/heuristic/severityrating.html
17. Hornbaek, K., Frokjaer, E.: Usability Inspection by Metaphors of Human Thinking Compared to Heuristic Evaluation. International Journal of Human-Computer Interaction 17, 357–374 (2004)
18. Landauer, T.K.: The Trouble With Computers: Usefulness, Usability and Productivity, pp. 219–221. M.I.T. Press, Boston (1995)
19. Spool, J.M., Perfetti, C., Brittan, D.: Designing for the scent of information. User Interface Engineering, Middletown, MA (2004)
20. Pirolli, P.: Information Foraging Theory. Oxford University Press, New York (2007)