Human-Machine Interaction Evaluation Framework

Hans Jander[1] and Jens Alfredson[2]

[1] Swedish Defence Research Agency (FOI), SE-164 90 Stockholm, Sweden
Hans.Jander@foi.se
[2] Saab AB, Aeronautics, SE-581 88 Linköping, Sweden
Jens.Alfredson@saabgroup.com

Abstract. The aim of the study reported in this paper was to use and evaluate a new methodological framework for Human-Machine Interaction (HMI) evaluation in system development for complex, high-risk and task-critical environments to assess overall HMI readiness. This has been conducted in the context of simulations in a state-of-the-art development simulator for fighter aircraft cockpit design in an industrial setting. The simulations included active and experienced military fighter pilots flying two civil navigational scenarios. The framework consists of already established evaluation methods and techniques combined with new influences inspired from risk management practices. A new HMI assessment survey has been developed and integrated into the framework. The results of the study are promising for the studied framework and also indicate some overlap when compared to existing practices regarding collected data. Applied within industry the framework can help leverage future HMI evaluations within system development.

Keywords: Usability, HCI, HMI, System evaluation, System Development.

1 Introduction

Within the domain of high risk and task critical environments there is a great need to incorporate end users iteratively in system development and design processes to be able to evaluate a suggested HMI-design in a relevant context (Hackos & Redish, 1998; Suchman, 2007; ISO 9241-210, 2010; Jander, Borgvall & Castor, 2011; Jander, Borgvall & Ramberg, 2012). This paper focuses on the evaluation step in the system development and design process. HMI-evaluations are not always prioritized and when evaluations are conducted the result from evaluations often comes in too late and suggested issues/improvements/changes in design are not always implemented due to time and budget constraints within projects. There are several reasons for this. One potential reason, that evaluations not always are integrated per default in the design process, is that there are no standardized evaluation procedures.

There is a need to develop evaluation methods that can be used, applied and adapted in system development and design to enhance overall system efficiency and meet the end user needs. Every millisecond that can be saved, every mental workload

M. Kurosu (Ed.): Human-Computer Interaction, Part I, HCII 2013, LNCS 8004, pp. 221–230, 2013.

decrease will improve the operator capability to perform their task in a faster, safer, and more accurate way.

Cost benefits aspects of using different evaluation methods needs to be considered before implementation within the industry.

This paper describes a study performed at Saab Aeronautics in PMSIM in Linköping, Sweden. PMSIM is a state-of-the-art development simulator for fighter aircraft cockpit design. The aim of this study was to evaluate a new methodological evaluation framework that has been developed within a research project in cooperation between Swedish Defence Research Agency (FOI), Saab Aeronautics, and Stockholm University. The project overall sponsor is the Swedish Governmental Agency for Innovation Systems (VINNOVA), within the National Aviation Engineering Program 5 (NFFP5). The focus of this study is not to evaluate the system that was tested, but rather to evaluate the developed the methodology.

The methodological evaluation framework developed in the project is further described in Jander, Borgvall, & Castor (2011), and Jander, Borgvall, & Ramberg (2012). The framework uses a variety of already established Human Factors (HF) and Human-Machine Interaction (HMI) evaluation methods and techniques such as think aloud protocol, mental workload measures, surveys and interviews combined with new influences inspired from risk management practices. A new HMI assessment survey has been developed and is integrated into the framework.

One of the new things within this methodological framework is the concept of use subjective weighting of parameters evaluated in the so called HMI assessment survey.

2 Objective

The overall objective of the reported study was to evaluate a new methodological framework for evaluating and assessing HMI in a fighter aircraft cockpit. Parameters investigated where:

- Time to perform evaluations
- Time for evaluation setup
- Time for analysis
- Type of data captured/collected
- "Know-how" needed to perform evaluation from the test leader perspective
- Test leader acceptance
- Test person (participant acceptance)
- Overall applicability of the methodological framework

3 Method

Two different evaluation methods approaches were used to evaluate characteristics of the systems HMI and was later compared. Method 1) New methodological evaluation framework; Method 2) A predefined survey addressing specific questions concerning

HMI functionality (benchmark). More specifically Method 1 was first used and was in the end complemented with Method 2.

Two test leaders conducted the evaluations. The evaluation was simulation based with three participants performing two missions in the flight simulator including the use of new functionality relating to HMI while performing predefined tasks using the system. On a meta-level an overall analysis were made to evaluate the two methods used to describe characteristics, e.g. pros and cons and give recommendations for future work.

3.1 Participants

All together five subjects, all male, participated in the study. Three were active or former fighter pilots from the Swedish Air Force and two persons with experience of system evaluation, one from Saab Aeronautics and one from the Swedish Defence Research Agency. The pilots were all classified as experienced fighter pilots with rudimentary experience in civil navigations procedures. The fighter pilots represented different experience levels. The first with approximately 8 years of working experience, the second with approximately 15 years of working experience, and the third with approximately 30 years of working experience. The two test leaders conducting the evaluation were both classified as experienced HMI-specialists, each with more than ten years of relevant working experience in the field. One was considerably more of an HMI generalist with expertise in HMI evaluation methods and the other was also considered as a specialist in the fighter aircraft domain. The two test leaders lead the evaluation procedure, but also in the end analyzed the result on a meta-level, e.g. describe method characteristics. Also, the role of an Air Traffic Controller (ATC) was used during the simulations to increase validity in the study.

3.2 Apparatus

The study was conducted in PMSIM (Display and Control Simulator) at Saab Aeronautics in Linköping. PMSIM is a state-of-the-art development simulator for fighter aircraft cockpit design. The simulator is a fixed base, dome simulator, where the visual surroundings are displayed on a dome with a radius of three meters, with a field of view of +/- 135 degrees azimuth and +90/-45 degrees evaluation. The simulated aircraft was a top-modern fighter aircraft.

3.3 Scenarios

Two pre-defined civil navigational scenarios was set up with the purpose of testing new system functionality to support pilots in civil navigation procedures including take off, holding, and landing. Especially new visual presentation of information regarding Area Navigation (RNAV) was displayed. Functionality and visual presentation regarding SID (Standard Instrument Departure) and STAR (Standard Terminal Arrival Route) were displayed, and the pilots interacted based on this information in the two scenarios.

3.4 Analysis

The interpretation of the results is made on a meta-level and is focused on the characteristics of the two different methods rather than the results of the specific system evaluation. More detailed descriptions of the analysis approach are described in the result section below.

3.5 Procedure

Each participant was given a short written description about the experiment, e.g. purpose, aim, and procedure. Then, each participant was presented and briefed about the new system for civil navigation procedures by a simulator instructor. Before entering the flight simulator cockpit the participants was informed how to use the Bedford rating scale for mental workload and how to think aloud when performing tasks in simulator.

Each participant performed two scenarios in the simulator using new system functionality and was asked to think aloud and highlight event-triggered events, and rate mental workload (MWL) according to the test leader instructions during the whole scenario. In average each participant were asked make MWL-ratings every fourth minute. Event-triggered comments and MWL-ratings were noted by the test leaders.

After completion of the simulation, participants were asked to report some spontaneous reactions and comments of the simulations and the new system used.

The participants were then asked to complete the HMI survey, facilitated by the test leader. They answered the survey by rating the importance of each of the 24 HMI criteria and rated the perceived criteria fulfillment of each criterion. The participants were also asked to make comments, give examples, make diagnoses on potential issues clarifying and motivating their choice of ratings. The ratings where based on the task performed and the system used in the simulator. This was explicit to the participants with the purpose of catching contextual aspects of use. An example of a criterion from the HMI survey is: Menus, symbols and texts are grouped in a logical way.

The participants were then asked to answer 8 questions survey regarding specific functions and displays of the system evaluated, also referred to as method 2. These questions were used as benchmark and comparison measure. An example of a question is: What are your comments on how data is presented on the center display?

There was no difference in the test procedure between the two evaluated methodological approaches except the tools used for data collection in the sense that both methods use fighter pilots as participants performing the same task scenario in the simulator.

In the end the participants were asked some questions regarding experiences of the overall applicability of the evaluation method and procedure just conducted.

All steps in the evaluation procedure was timed, think aloud and event triggered comments and MWL was noted by the test leaders. The test leaders were using predefined test protocols.

After the system evaluation sessions with the three fighter pilots, data was analyzed.

4 Results and Analysis

4.1 Time to Perform Evaluation

The average time to conduct on evaluation was 3 hours and 25 minutes. Some more time was needed (in average 40 min) to perform the HMI-survey (method 1) compared with the benchmark evaluation survey (method 2).

4.2 Time for Preparation

The preparation time for the evaluation is very dependent on the apparatus and test scenarios needed and personnel involved. Test scenarios already existed and the simulator was up and running. The total preparation time for the evaluation time is estimated to 3 working days for the evaluation team. If new test scenarios needs to be designed more time is needed.

4.3 Time for Analysis

The results collected from the benchmark evaluation survey are relatively straight forward and easy to interpret due to the design of the specific questions. Most of the answers referred mostly to describing and guiding specific system characteristics. The results from the HMI evaluation framework require more time for analysis. There are many more dimensions of the HMI that are investigated and the results from MWL-ratings, event-triggered events, and HMI-survey all needs in depth analysis that are further described below. The results from the HMI evaluation framework is not only describing and guiding specific system characteristics, but also describes more general system characteristics complemented with potential prioritizing of identified issues (as described under the section 4.10 **"**Comparison of data from HMI assessment survey and baseline survey").

The time for analysis of the results from the HMI evaluation framework is approximately 1 day per participant and 1-2 hours per participant for the benchmark evaluation.

4.4 Mental Workload Measures

Bedford rating-scale were used to rate Mental Workload (Castor, 2009). The scale consists of ten steps (1=very low MWL and 10 very high MWL). See Table 1 for the three pilot participants' MWL-ratings.

Due to the lack of a control group performing the test scenarios without using the new system to support civilian procedures at take-off, holding, and landing it is hard to make any conclusions how the evaluated system specifically affected MWL. In a few cases MWL-rating were high but considering participants additional comments (think aloud) these MWL-ratings cannot directly be deduced to this specifically system functionality, rather to overall system functionality (which is an interesting finding) and different participants experience levels.

Table 1. Mental Workload ratings

Participant (P)/Scenario (S)	Number of ratings	Mean	Standard Deviation
P1/S1	10	4.2	1.5
P1/S2	8	4.9	2.2
P2/S1	14	4.2	1.2
P2/S2	9	5.2	0.8
P3/S2	10	4.6	1.4
P3/S2	9	4.3	0.9

4.5 Think Aloud Event Triggered Events

Only a few relevant event triggered comments referring to system characteristics were articulated during the test scenarios. Due to the relatively non-complex tasks and low dynamics in the scenario, very few frustrations or other events were highlighted. A few times the participants raised questions how to navigate in system menus. Also some comments were made that referred to specific design solutions and suggestions regarding the interface.

4.6 HMI-Survey

The participants experienced some redundancy between some criteria in the HMI-survey. For example, the criterion statement "The system empowers me to complete the assigned task in the best possible way" is similar to the criteria statement "I feel that the system fulfills my needs". Overall, all criteria were rated as important on the six-grade rating scale. This indicates that almost all criteria in the survey were considered relevant for the system tested in this specific context with very few exceptions.

4.7 Participant´s Comments and Justifications on HMI-Ratings

The rated criteria value and the rated criteria fulfillment value was complemented with comments with the purpose of motivating, clarifying, and justifying ratings. An example was when one participant rated the criterion "I have a feeling of achieving high task effectiveness when using the system" as 4 (rather important) on the importance scale and as 2 (almost totally fulfilled) on the fulfillment scale. An additional comment made by the participant on the rating was; "I prefer accuracy prior to efficiency in the context of civil navigation". This example illustrate that the importance of the different criteria might differ in another context and this aspect is captured in the evaluation framework.

The comments made by the participant added great value and meaning to the criteria ratings in the survey. In some cases spontaneous design issues were addressed and some specific design suggestions were articulated.

4.8 HMI Assessment Matrix (Analysis Tool)

The product of the rated criteria value and the rated criteria fulfillment value from the HMI-survey resulted in a number from 1-36. Low numbers was assumed to indicate that there are no design issues, e.g. HMI is ok. High numbers indicate that there are some design issues that needs to be considered, e.g. HMI is not ok. Though, the result of the study indicates that it is very hard to draw any conclusions from just a number from 1-36. There are several reasons for that. For example, if two criteria have the same product value it is hard to choose which of them is the most important to consider. Also, in some cases in the study the product value was relatively high but considered additional comment made by the test person indicated that there actually was no issue. Therefore, it is very important to consider the column of comments made by the test person for each of the criteria. The result of using the HMI-matrix as an analysis tool shows that it is just a complement to other collected quantitative data. The quantitative data gives power to the qualitative data and the qualitative data dress the quantitative data with meaning. To give a meaning and make conclusions of just a number between 1-36 alone is in this case inappropriate and even hazardous.

4.9 Benchmark Survey

The benchmark survey (method 2) consisted of eight specific questions regarding the functionality of the tested system. Some of the questions were not answered by the participant due to that they did not use all the functions that the questions addressed. In general, given answers addressed specifically system characteristics.

4.10 Comparison of Data from HMI Assessment Survey (Method 1) and Baseline Survey (Method 2)

In order to compare the results of the data collected from comments made in the HMI-survey with the answers from the benchmark survey, a taxonomy was created to classify comments from the HMI-survey and answers from the benchmark survey. Four classes were created (see table 2). Class 1, 2, A, and B: were class 1 refer to comments and answers on general system characteristics; and class 2 refer to comments and answers on specific system characteristics; and class A refer to describing comments and answers; and B refer to guiding comments and answers.

Table 2. Taxonomy used for analyzing results of the HMI-survey (method 1) and benchmark survey (method 2)

Class	1	2
A	Describing general system characteristics	Describing specific system characteristics
B	Guiding general system characteristics	Guiding specific system characteristics

When comparing the results from the HMI-survey and the benchmark survey it is obvious that most of the answers referring to guiding and describing specific system characteristics was collected from the benchmark survey but the comments from the HMI-survey also give some guidance regarding specific system characteristics. On the other hand, the HMI-survey also describes and gives guidance on specific system characteristics and also describing general system characteristics. When conducting system evaluation specific functionality is hard to isolate from the overall system and this is probably not always even desirable. There were also some overlap and redundancy in answers between the HMI-survey comments and the benchmark-survey answers.

4.11 Know-How Needed to Perform Evaluation from the Test Leader Perspective

To be able to interpret result accurately it was vital to have at least on test leader with domain experience. It also leverages the credibility in the relation with the participants. For practical reasons it also helps with experimental setup and administration to have some "inside" the organization were the evaluation will take place. The know-how needed could also consider the three different stages when conducting system evaluation: 1) Preparation; 2) Performing; 3) Analyzing. For preparation someone from the organization were the evaluation will take place is vital to make necessary arrangements (scenario design, simulator set up including simulator operator/s). Some domain expertise is needed to design questions referring to this study benchmark test. For preparation of test protocol of the HMI assessment framework, domain expertise is not necessarily needed. When performing the evaluation two test leaders are needed. At least one should have domain expertise and at least one should have experience of HMI-evaluations. During the analysis it is desirable to include the test leaders who have conducted the evaluation with the motivation of capture details during the evaluation in order to transform the result to valid conclusions and communicate to the design team.

4.12 Participants' (Pilots and Test Leaders) Acceptance

Both the test leaders and the pilots experienced positive acceptance of the new methodological evaluation framework and judged the framework as relevant, valid and easy to conduct.

5 Conclusions

The study shows promising for the studied HMI evaluation framework and also indicated a few overlaps with existing practices within the industry regarding results in identification of specific describing and guiding system characteristics data. The HMI evaluation framework also identified more general system characteristics data, referring to the whole system used, not only the evaluated system tested in isolation.

The use of a combination of qualitative (survey comments, think aloud, and interview) and quantitative (survey and MWL ratings) measures suggested in the new framework will leverage HMI-evaluations and help system designers to find, describe and prioritize potential design issues into further design iterations. Additional comments on each criterion are vital to consider before making conclusions of numerical values in isolation. More studies needs to be conducted to validate the applicability of the suggested evaluation framework evaluating other systems in different contexts within the studied domain. The studied framework can both be used for benchmark and acceptance tests, but also for formative and diagnostic testing. The frameworks ability of considering contextual aspects and the combination of using both quantitative and qualitative data gives considering advantages.

6 Discussion and Future Research

The new methodological evaluation framework approach (method 1) investigated in this study shows promising results in system evaluation. Some of its advantages are the explicit use of the concept of weighting which is rather new in systems evaluation, even though the use of weighting sometimes is used more implicitly in evaluations. One way of catching the right context of use of a system in evaluations is the assumption that the importance of identified HMI criteria might differ between different systems, tasks, and users (Frokjaer, Hertzum, & Hornbaek, 2000). The use of weighting considers these aspects and gives valid results in evaluations. The evaluated methodological evaluation framework is generic and can be used for evaluating a variety of systems within the domain.

One potential problem using specific questions (method 2) about system functionality is that the answers tend to be quite isolated and just relate to the specific system tested. In a complex system HMI like a fighter aircraft cockpit there is always other interactions needed that relate to other overall system functionality as well. Therefore, there is a need to conduct systems evaluation using the new functionality integrated with the overall system in a relevant scenario to capture the right context of use. However, the use of specifically addressed questions can on the other hand give valuable insights about specific system characteristics and these questions can serve as a complement to the methodological evaluation framework.

Most of the HMI criteria were rated as important and that might lead to problems when identifying design issues when using the HMI assessment matrix alone referring how to in the best way prioritize identified issues in further iterations. Therefore additional comments need to be carefully considered during the analysis.

The study setup and experiment design could use a control group performing the same task without using the system new system in order to make comparisons regarding how the new system affected ratings and comments. However, this was not possible due to lack of participants and time constraints.

In this case the evaluated system was not very complex and the task performed in the simulator was relatively simple. A new study is needed to evaluate another system, preferably in a highly dynamic scenario with increased task complexity to further evaluate the new methodological approach.

During the analysis phase in this study no end users (i.e., the pilots) received feedback and were consulted to validate the test results. Due to lack of access of the participated pilots in the analysis phase this was not done. It would have been recommended to consult the end users for a double check to make sure the results and analysis is valid, also from the end users perspective before writing final test report.

The benchmark survey (method 2) used in this study requires some extra time to prepare compared with the survey used in the evaluation framework. The benchmark survey also addresses very specific system characteristics and sometimes missed to catch more general system characteristics that were identified using the evaluation framework.

The evaluated framework puts focus on both finding pros and cons regarding system characteristics. The classification/taxonomy described just describes system characteristics in four dimensions. However, each of the system characteristics could also describe the identification of both positive and negative aspects of the system. Traditionally HMI-evaluations are primarily concerned with identifying problems, while both negative and positive system characteristics were identified in this study. This is of great importance to designers who also needs to know what the systems strengths are.

References

1. Castor, M.: The use of structural equation modeling to describe the effect of operator functional state on air-to-air engagement outcomes. Doctorial Thesis No. 1251, Linköping University, SE-581 83 Sweden (2009)
2. Frokjaer, E., Hertzum, M., Hornbaek, K.: Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In: Conference on Human Factors in Computing Systems (CHI 2000), N.Y., April 1-6 (2000)
3. Hackos, J., Redish, J.: User and task analysis for interface design. Wiley, New York (1998)
4. ISO 9241-210:2010. Ergonomics of human-system interaction, Part 210, Human-centred design for interactive systems, Geneva, Switzerland (2010)
5. Jander, H., Borgvall, J., Castor, M.: Brain Budget- Evaluation of Human Machine Interaction in system Development for High Risk and Task Critical Environments (FOI-R–3272–SE) (2011)
6. Jander, H., Borgvall, J., Ramberg, R.: Towards a Methodological Framework for HMI Readiness Evaluation. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56 (2012)
7. Suchman, L.: Human-Machine Reconfigurations, Plans and Situated Actions. Cambridge University Press, NY (2007)