A Validation Approach for Complex NextGen Air Traffic Control Human Performance Models

Brian F. Gore[1] and Paul Milgram[2]

[1] San Jose State University/NASA Ames, MS 262-4, PO Box 1, Moffett Field, CA, USA
Brian.F.Gore@nasa.gov
[2] Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, Canada
milgram@mie.utoronto.ca

Abstract. Validation is critically important when human performance models are used to predict the effect of future system designs on human performance. A model of air traffic control (ATC) operations was validated using a rigorous iterative model validation process that illustrated the success of representing ATC operations in NextGen en route operations. A gold-standard model was compared to three model iterations that represented different task management and human time estimation processes when dealing with handoff operations.

Keywords: Human performance model validation, air traffic control, NextGen.

1 Introduction

Human performance modeling is the process whereby human characteristics are embedded within a computer software structure that represents a simulated human operator interacting with a simulated operating environment. Integrated human performance models (HPMs) simulate and predict emergent behavior based on multiple, interacting sub-models of human behavior, such as perception, attention, working memory, long-term memory and decision-making. This is accomplished typically by incorporating sub-models of different aspects of human performance that feed both forward and back to other constituent models within the human information processing system. The use of appropriate and validated integrated HPMs can support the basic human factors principle of predicting the impact of alternative design options early in the system design process. Such HPMs may also be used synergistically with human-in-the-loop (HITL) studies, especially during the development of complex systems, a time when events cannot be studied fully with HITL subjects due to safety concerns, cost considerations, or practical difficulties associated with simulating very rare events.

Complex systems are those that include human operators interacting with actual technology and automation, to carry out multiple interacting, and often conflicting, tasks. These systems often involve time-critical tasks, that is, tasks that typically have a specific onset time and a specific time by which the task needs to be completed. Together these define a window of opportunity for the action to take place. For such systems, the dynamic interactions among system elements often form critical couplings for control of the system by the human.

V.G. Duffy (Ed.): DHM/HCII 2013, Part I, LNCS 8025, pp. 28–37, 2013.

One of the most significant hurdles facing modelers is the challenge of validating these integrated HPMs, a goal without which the credibility of any model predictions will clearly be greatly reduced [1]. Most validation efforts to date have been in the area of simpler engineering models and cognitive architectures [2], with only a small number of attempts to validate integrated HPMs. Furthermore, of the validation efforts that have been conducted for integrated models, there is little agreement as to what constitutes appropriate validation techniques and measures [1]. The development of these integrated HPMs is in its infancy, and so too are the validation techniques. There is thus a real need for the advancement of techniques and approaches for validating complex models.

2 Validating Complex Models

Modeling human behavior in complex systems such as air traffic control is very complicated, particularly when the human's tasks are highly cognitive in nature and they interact in a closed-loop fashion with other operators and environmental factors. Since cognitive tasks are not directly observable, it is very difficult to objectively validate such complex models. As a field, our ability to model these complex tasks and demonstrate that such models of human behavior validly represent actual human behavior is in its infancy. Many HPM validation efforts often rely only on subjective or qualitative measures, as opposed to objective, quantitative measures. Thus, one major objective of our work is to focus on quantitative validation techniques that can be used to demonstrate that a particular model represents human cognitive processes. The present research highlights a method that was followed to develop valid HPMs of time management in a complex operational environment exemplified in an air traffic control domain.

3 Defining Model Verification and Validation

Model verification and validation are essential elements of any modeling effort. M*odel verification* is the process of determining whether a simulation model and its associated data behave as intended by the model developer / analyst. *Model validation* is the process of determining the degree to which a model or simulation and its associated predictions are an accurate representation of the real world, from the perspective of the intended users of the model or simulation [3]. Both model verification and model validation must be considered when attempts are made to validate a model, particularly as models increase in complexity.

Model validation can take many forms, ranging from common qualitative approaches to quantitative approaches [4]. For the purpose of brevity, only the quantitative approach will be focused on in the current article. Obtaining a quantitative measure of the similarity between a model's behavior and empirically determined human behavior is a complement to the qualitative approach. More explicitly, a quantitative test for a model's validity is the degree to which the model's output resembles the behavior that would be expected from the real world. Quantitative approaches are

traditionally statistical in nature and attempt to measure the degree to which a model's data are similar to an empirically collected set of data. The recommended statistical tests used to measure the similarity between the data sets include goodness-of-fit tests (r^2) to assess trend consistency; ANOVAs to compare human and model data sets; root mean squared scaled deviations to assess the exactness of matching; and chi-square analyses to assess whether the underlying distributions of the two data sets (model, real world) can be regarded as coming from the same population [4].

Graphical comparisons are also an effective model validation approach, particularly as a first validation phase for initial testing of model performance [3]. Using that approach, the graphs of values of model variables over time are compared with the graphs of values of system variables, to investigate, for example, similarities in periodicities, skewness, number and location of inflection points, logarithmic rise and linearity, phase shift, trend lines, or exponential growth constants. The histogram is an estimate of the density function and is another effective graphical technique, for examining data symmetry, skewness and kurtosis.

4 Simulated Environment and Validation Approach

A complex, time-critical environment, namely the ATC environment, was used as a test-bed to develop and exercise validation techniques that concentrate on validating the time-relevant aspects of the model. Using an iterative develop-validate process, the test-bed model was augmented with sub-models (embedded models) that represent the processes required to execute a series of procedures [2], in this case, time management procedures.

Three human behavioral components occur in a time management environment; i) task management, ii) time estimation, and iii) time management. (Because the sub-models can be modified individually, any differences in model output can be attributed to the sub-model under investigation.) The particular sub-models used were based on a synthesis of existing literature on the manner in which humans' time management (task management and time estimation) changes in the face of dynamics of the operational environment, as a function of time pressure, and thus of perceived workload.

Validating the time management model required three steps. Starting with an appropriate baseline model, the first step is to assess the *task management* portion of the model. The second step is to assess the *time estimation* aspect of the model. The third step is to assess the *time management* aspect of the model. (Such an approach can be extended to any complex domain.)

The domain that was chosen was from a FAA HITL simulation of the Future En-route Workstation Study (FEWS) [5]. This dataset included three workload levels: low, medium, and high, as defined by the number of aircraft travelling in a generic airspace and the presence of assistive technologies (datalink - DL). Air Traffic Controllers (ATCos) were required to schedule tasks that differed in

priority, including conflict resolution, aircraft hand-offs between sectors, and routine communications.

The baseline (BL) HPM of the ATCo, programmed in Micro Saint Sharp was used as the *gold-standard,* against which further model augmentations would be compared, as this model had been deemed by others to validly represent the FEWS performance [6]. The baseline model, depicted in Figure 1, assumed that ATCos implement nominally optimal strategic task scheduling, and that the ATCos' estimate of time passage is perfect and with no effect on workload.

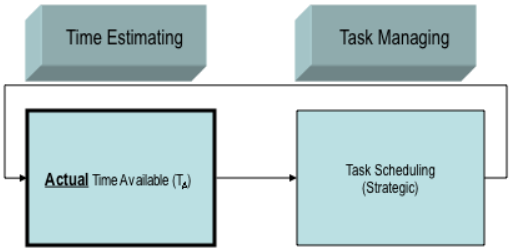


Fig. 1. Baseline Model of Task Management Behavior

To verify and validate this BL model, performance on several output variables from the model was compared to the human ATCo performance collected during the FEWS simulation. The non-human portion of the model was verified in terms of environmental performance (the number of aircraft travelling in the airspace, the flight plans, etc.). Modeled ATCo workload and queue length were assessed to verify that the model behaved as expected – that is that ATCo workload and the number of items in the queue increased as a function of task load, in the same manner as occurred in the FEWS HITL data.

A large contribution of the present work extends the validated environmental model to the modeled operator behaviors. The ATCos' Receive Handoff Duration (RHD) data – the elapsed time between the first moment at which a particular aircraft could have been handed off and the moment at which it actually was handed off – were used as the main validation measure.

The RHD was first explored at an aggregate mean level and compared by both t-tests and chi-square analysis to the FEWS RHD data. Figure 2 illustrates that no significant difference was found in the mean RHD times for the low workload between the FEWS and the baseline model, $p>.05$. A significant difference did exist, however, for the medium workload between the FEWS data and the baseline model predictions, $t(99) = 2.07$, $p<.05$. The same was true for the high workload, $t(80) = 5.51$, $p<.001$. Using these measures, validation of the BL model was supported in the low workload conditions, on the basis of not having detected a significant difference between model and FEWS data ($p>.05$). Validation was not supported in the medium and high workload condition, however ($p>.05$).

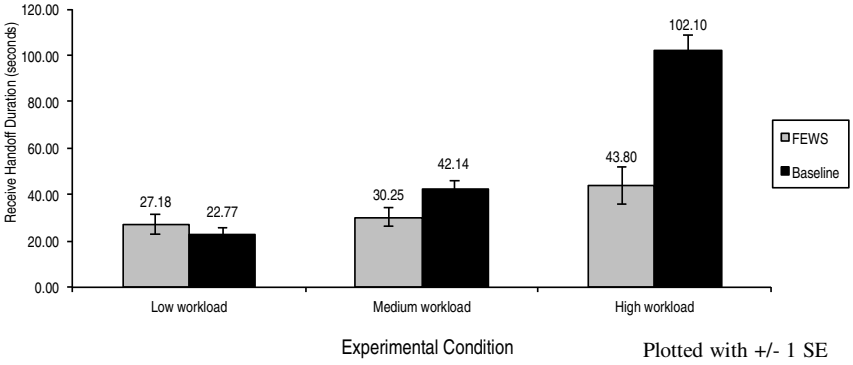


Fig. 2. Average RHD Times from the FEWS data and Baseline Model

A novel approach, dubbed *Time Correspondence (TC) graphs*, was implemented, to compare the BL model with the FEWS data. In each TC graph, such as Figure 3, the top horizontal axis shows the timeline of occurrence of FEWS events, while the lower axis shows events modeled by the HPM. The lines joining corresponding events are perfectly vertical when the times were the same in both FEWS and model. Lines that slant down to the right indicate that model times occurred later than corresponding FEWS times, and lines slanting down to the left indicate that model times preceded corresponding FEWS times. The ensemble of all lines resulting from a simulation thus form a holistic visual indication of the goodness of model fit.

The BL model results in Figure 3 show the window open times on the left and the window close times on the right, since it is necessary to look at both in order to correctly attribute the reason for any difference in the RHD times. In the figure we see that the BL model was operating too liberally, allowing the possibility of 'too many' tasks to enter into a queue - which produced delayed processing of some tasks and thus later task onset predictions than those of the real operators. The TC graphs, supported by Spearman correlation coefficients, also showed that the order in which tasks were conducted differed between the model and the FEWS simulation.

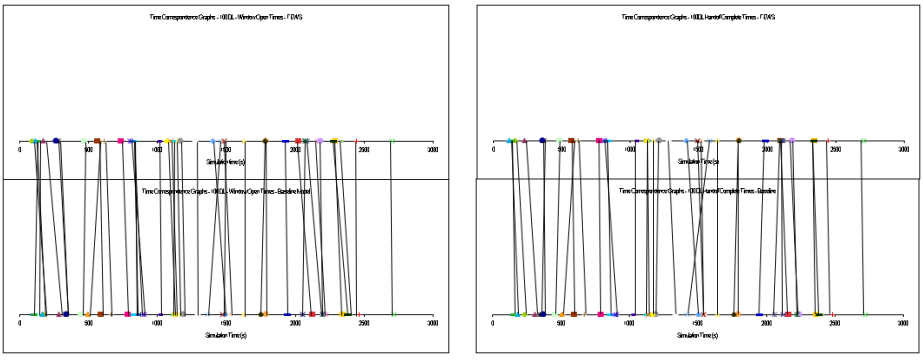


Fig. 3. TC Graphs - Window Open and Close Times (Baseline Model)

Two explanations for the breakdown in the BL model performance are offered. First, the strategy that the ATCo used to process the handoff tasks differed between the FEWS and the BL model. Second, it could be that the human operators simply failed to estimate the passage of time accurately due to excessive workload [7], as the model assumed that the operator always has "perfect" awareness of available time to complete a task and, furthermore, it assumes that tasks are completed quickly, in the right order and at the right time. Underestimating time available results in tasks being scheduled for completion early within the window resulting in a burst of early responses. That model modification is described after the task management section.

5 Development of a Framework of Time Management

The Time Management framework that follows was used to guide the development of two time management information processing models – time estimation and task management [8] - which are to be called whenever the human operator engages in a time sensitive task. As shown in Figure 4, the time management framework includes a workload projecting component, a time estimating component, and a task-managing component.

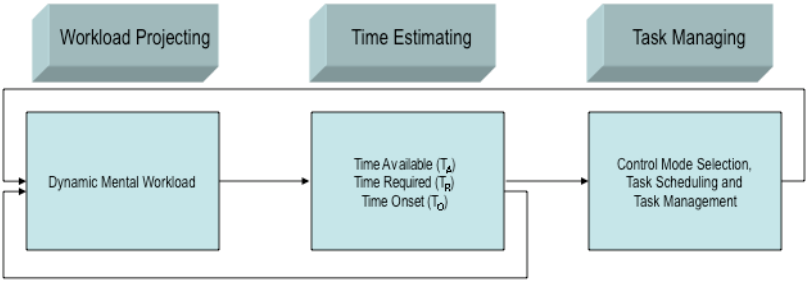
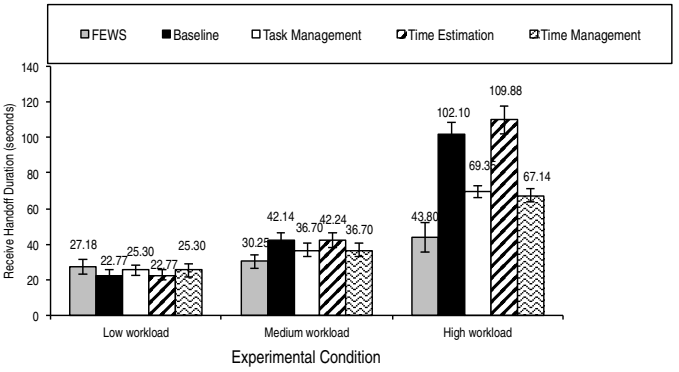


Fig. 4. Workload Projecting, Time Estimating and Task Management framework

5.1 Baseline (BL) Model with Task Management Modification

The BL model was augmented to account for the change from *strategic* to *opportunistic* control that occurs in high-workload tasks, as indicated in the literature [9]. The *task management (TM)* model utilized a 'conservative bias' paradigm, where the human operators are expected to complete all tasks in the order in which they are encountered, thereby shifting performance times towards "early" responses within the window of opportunity for certain high priority tasks. The right-most box in Figure 4 has been provided with a feedback loop around itself. This serves to impact the task ordering and onset times using the *opportunistic* control mode in the high workload condition. This new model does not allow multiple tasks to collect in the task queue; it forces the operator to manage tasks *opportunistically*, rather that strategically. That is, planning is limited and the environment drives decisions.

Using the same validation measures as for the BL model, it was apparent that limiting the queue length to zero made the model perform more "conservatively" and succeeded in bringing the modeled RHD times closer to the FEWS RHD times, with no significant difference between the FEWS and the low and medium workload conditions (see Fig 5; p>.05). Similar to the BL model, the RHD times in this TM model were significantly shorter than in the original baseline model in the high workload condition (although still significantly delayed relative to FEWS (t(80)=3.30, p<.05).



Plotted with +/- 1 SE

Fig. 5. Average RHD Times from FEWS and the Four Model Iterations

5.2 Baseline Model with Time Estimation Modification

The Time Estimating Box (TE; center box) in Figure 4 was modified to account for the degradations of human estimates of time passage as a function of workload. A quantitative verification effort of the TE model was conducted using a simple but non-trivial ATC task network model. This "generic computational model" was run 10000 times in two scenarios (baseline and TE model) at each of 5 workload levels (low, low-medium, medium, medium-high, and high) to verify the impact of time misestimates (time error) on task scheduling within the model. This verification phase revealed that the baseline model performs very close to zero mean error, as expected (Figure 5). The time estimation model time error output, on the other hand, revealed that the aircraft are being descended at increasingly late times as workload increased.

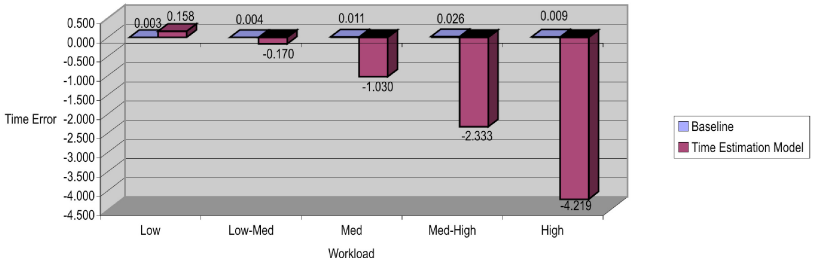


Fig. 6. Mean Time Error of 10000 Generic Model Runs to Verify TE Model Performance

As illustrated in Figure 6, the baseline model performed very close to optimally, with all of the time error values in the positive direction, meaning that the aircraft were descended early. It also illustrates the non-linear relationship that exists between workload and TE as measured by the time error in ATCo descending an aircraft.

As a result, a formal validation effort of the TE model in a specific ATCo environmental context was conducted to determine the generalizability of the TE model. The BL model's strategic task scheduling was combined with the TE underestimates of time passage as a function of increasing workload. The same validation approach that was applied to the BL model development iterations was applied to the TE model iteration. As can be seen in Figure 5, the TE model did not improve RHD predictions as compared to the baseline model. Relative to the FEWS data, the RHD times produced by the TE model were not significantly different from those produced from FEWS in the low workload condition ($p>.05$), but they were significantly higher in the medium ($t(99)=2.08$, $p<.05$) and high workload conditions ($t(80)=5.85$, $p<.05$).

5.3 Baseline Model with Time Management (TM+TE) Modification

To perform the *Time Management* (TM+TE) augmentation, the Time Estimating (center) and the Task Management (right most) boxes in Figure 4 were modified to account for the task management and the human estimates of time passage degradations that occur as a function of workload. Using the validation approach developed in this effort, it was apparent that limiting the queue length to zero in the TM model made it perform more "conservatively" and including a time estimation model succeeded in bringing the modeled RHD times closer to the FEWS RHD times than any of the other models alone. Relative to the FEWS data, the RHD times produced by the *Time Management* model were not significantly different from those produced from FEWS in the low ($p>.05$), and the medium workload conditions ($p>.05$) but they were significantly greater in the high workload conditions ($t(80)=2.28$, $p<.05$) (Figure 5).

The TC graphs also show that the order in which tasks were conducted in the *Time Management* model differed between the model and the FEWS simulation as illustrated in Figure 7 for the window open (left) and window close times (right).

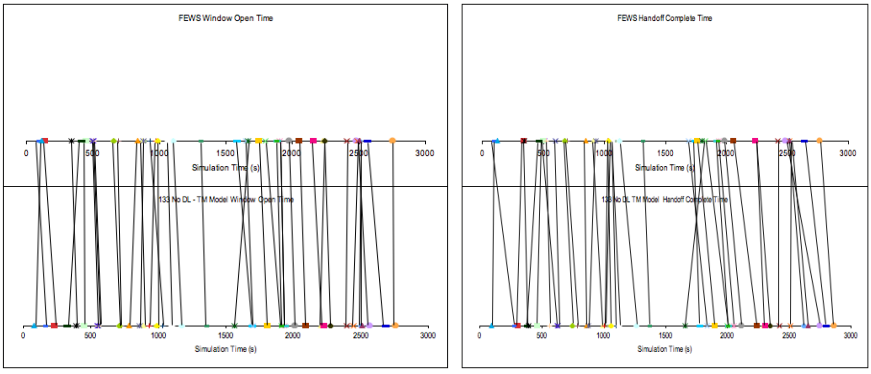


Fig. 7. TC Graphs - Window Open and Close Times (High Workload Model)

The TC window open time graphs illustrate that the order and the onset times are not precisely the same. While the overall data suggest that the high workload condition produced the greatest difference in RHD times, this effect can not be attributed primarily to the time that the window opened, since the present analysis suggests that the number of window open times that differed in order from the modeled and actual aircraft were fairly consistent across traffic conditions. The same holds true for the window close times.

It appears that the *Time Management* model did not succeed in bringing the model's RHD performance closer to the FEWS than the baseline model alone. Both the baseline and the *Time Management* model predictions of RHD remain significantly different than those produced by the ATCos in the FEWS experiment for the high workload condition. There did not appear to be any added benefit to the RHD prediction by both TM and TE together, although the *Time Management* manipulation did bring the mean RHD times closer than any of the other model manipulations alone.

6 Discussion

The recent proliferation of human-system models has resulted in highly complex human behavior models being used to generate predictions of operator performance within increasingly complex operational domains (e.g. process control, aircraft, and ATC operations, etc.). This proliferation is certain to continue along its growth path in the foreseeable future as computer technologies increase and the software implements more accurate representations of the human-system relationship. Many of the models that have been developed for system predictions have undergone some degree of verification and validation. However, creating valid behavioral models of a human is a challenging endeavor, particularly because of the complexity of human behaviors, which are further heightened when integrating multiple models that comprise the system. Assumptions made for one sub-model may interact with other sub-models and may invalidate the system prediction. As a result, it is vital that the complex human models that are used to generate predictions of human-system performance be designed and validated in accordance with a principled approach.

Validating the model using a limited number of validation measures (often only one measure) allows model developers flexibility with respect to the manipulation that will be made to the model to get it to perform consistently with the input data. It is often quite easy to tweak a model to perform well on one measure, while sacrificing the validity of other measures. When model analysts change a model's parameters, they typically do not examine the performance of the integrated representation of the model; rather they look at the effect of the individual parameter that they tweaked. While this is arguably an appropriate validation process for some small, non-integrated models, it is advisable that the more integrated and closed-loop HPMs conduct validation efforts use multiple human performance measures.

In summary, this research has introduced and demonstrated a comprehensive iterative develop-validate *approach* for validating a complex, closed-loop model of air traffic control using multiple measures at varying levels of fidelity designed to

provide a validation approach for *time-sensitive* tasks. A series of objective and quantitative validation measures were applied to assess the validity of a baseline model that was then carried through as model iterations were completed. The iterative approach enabled the assessment of the impact of each model manipulation to determine whether the model developed operated verifiably and validly. It is only with such a rigorous approach that the models that are developed for complex human-system operations can be deemed credible representations of actual human performance.

Acknowledgments. We would like to express our sincere appreciation to Professors Daniel Frances, Mark Chignell, Baris Balcioglu, to Dr. Ronald Laughery, and to Mr. Ken Leiden for their input and guidance throughout this research project. This research was conducted as part of the first author's doctoral dissertation at the University of Toronto in the Department of Mechanical and Industrial Engineering.

References

1. DMSO 2001 Defense Modeling & Simulation Office (DMSO): Verification, Validation, And Accreditation (VV&A) Recommended Practices Guide (RPG): Special Topic - Validation of Human Behavior Representations (Website) (September 25, 2001), http://www.msiac.dmso.mil/vva/Special_topics/hbr-Validation/default.htm
2. Baron, S., Kruser, D.S., Huey, B.M.: Quantitative modeling of human performance in complex, dynamic systems. National Research Council Washington DC Panel on Human Performance Modeling, Washington (1990)
3. Balci, O.: Verification, Validation, and Testing Techniques. In: Banks, J. (ed.) Handbook of Simulation: Principles, Methodology, Advances, Applications, And Practice, pp. 335–427. Wiley & Sons, Inc., N.Y. (1998)
4. Campbell, G.E., Bolton, A.E.: HBR Validation: Interpreting Lessons Learned From Multiple Academic Disciplines, Applied Communities, And The AMBR Project. In: Gluck, K.A., Pew, R.W. (eds.) Modeling Human Behavior With Integrated Cognitive Architectures: Comparison, Evaluation And Validation, pp. 365–395. Lawrence Erlbaum & Associates, New Jersey (2005)
5. Willems, B.: Future En Route Workstation (FEWS) Study. FAA William Hughes Technical Center Atlantic City International Airport, New Jersey (2005)
6. Leiden, K., Kamienski, J.: DAG CE-6 Modeling and Simulation Studies. DAC Program Review Presentation. NASA Ames Research Center, Moffett Field (2006)
7. Hart, S.G.: The Prediction And Measurement Of Mental Workload During Space Operations. In: NASA Space Life Sciences Symposium, National Aeronautics and Space Administration, Washington, DC (1987)
8. Gore, B.F., Milgram, P.: The Conceptual Development Of A Time Estimation Model To Predict Human Performance In Complex Environments. In: Ninth Proceedings of the Annual SAE International Conference and Exposition - Digital Human Modeling for Design and Engineering Conference, SAE Paper # 2006-01-2344. SAE, Inc., Warrendale (2006)
9. Hollnagel, E.: Cognitive reliability and error analysis method (CREAM). The Alden Group, Elsevier Science, Oxford, UK (1998)