Quantitative Scoring System to Assess Performance in Experimental Environments

Ron Becker, Sophie-Marie Stasch, Alina Schmitz-Hübsch, Sven Fuchs
Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE
53343 Wachtberg, Germany
Email: {ron.becker, sophie-marie.stasch, alina.schmitz-huebsch, sven.fuchs}@fkie.fraunhofer.de

Abstract – **A quantitative scoring mechanism based on signal detection theory was developed in the context of an experimental command-and-control environment. The scoring approach was designed to include well-established evaluation criteria of performance metrics and to enable insights into various cognitive and behavioral processes of the subjects. Cognitive processes on a perceptual, sensory, and motor level were linked to subtasks similar to the warship commander task. Signal detection theory provides a theoretical rationale for the quantitative scoring mechanism. Due to the generalizability of the scoring approach, a flexible application to a wide range of experimental tasks should be possible. Considerations and lessons learned are discussed.**

Keywords – Behavioral processes; Cognitive processes; Human machine interaction; Command-and-control; Quantitative performance metrics, signal detection theory

I. INTRODUCTION

Measuring human performance in complex experimental tasks can be challenging. Generally, two approaches are available to measure user performance in such experiments: Qualitative performance assessment aims at gaining an in-depth understanding of the matter of interest as well as its context. Since the observer's point of view is internal, the results tend to be subjective and difficult to verify. Consequently, conclusions based on qualitative assessments may not be replicable. In contrast, quantitative approaches objectify performance assessment and enable the use of inferential statistics, such as significance testing [1][2]. Quantitative performance assessment aims at measuring human performance through predefined metrics that can be calculated in an automatic manner, resulting in a numeric value, such as a score. Moreover, it allows for flexible adjustment of the scoring mechanism to reflect specific characteristics of an experimental task (e.g., task priorities). This is important, given that adequate scores can reveal underlying cognitive processes involved in task performance. For instance, by combining time and the amount of errors to complete a given task, one can derive insights about the speed-accuracy tradeoff of the participant.

Ducheneaut, Moore and Nickell [3] provide an example of how a quantitative assessment method can reveal deeper insights into complex behavioral processes than a qualitative assessment method could. While exploring the concept of sociability in massive multiplayer online games, the authors gained better understanding of the matter by investigating the number, frequency, and length of visits in social places. However the qualitative results could not reveal how generic and how representative the observed activities were, so they turned to quantitative analysis. Other experimental testbeds benefit from a quantitative performance assessment in the same way, for example the Warship Commander Task (WCT) [4].

Safety-critical vigilance tasks can be found in many domains, such as air traffic control, driving on highways, or in the control room of nuclear power plants. The Warship Commander Task is one example of such a task. With the primary goal of protecting their own ship from hostile tracks appearing on a simulated radar-screen, WCT users must complete multiple subtasks of different priorities within limited time. These hierarchically organized subtasks include identifying all unknown tracks, as well as warning and engaging identified hostile tracks. The engagement of a hostile track can only be performed after its warning. Similarly, the warning of a hostile track can only be performed after its identification (see Figure 1), resulting in a hierarchical task structure.

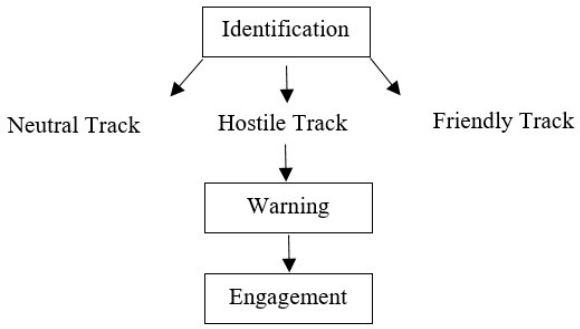

Figure 1. Subtasks of the WCT

Cognitive processes involved in executing a subtask in the WCT are identified based on the human processor model developed by Card, Moran and Newell [5]. The human processor model describes the calculation of reaction times based on the time needed for perceptual, cognitive, and motor processes. Figure 2 illustrates the cognitive processes involved in executing a subtask of the WCT: Visual attention and perception are a necessary prerequisite for detecting an object and discriminating it from the background of the screen. After successful detection of a relevant object, a rule-based decision is necessary to identify the object and determine if further action is necessary. In case of the WCT, such a rule can be found in the warning and engagement subtasks. The operator must decide if a track has to be warned or engaged based on its identity and distance.

 ISBN: 978-1-61208-870-9

The decision-making process is followed by task execution. The physiomotor response of clicking the button representing the required action concludes the subtask. For the identification subtask, the operator clicks on a button named “IFF” (“identify friend foe”). For executing a warning or an engagement, the operator clicks on a button named “warning” or “engagement”.

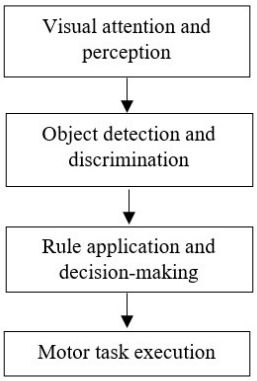

Figure 2. Cognitive processes involved in WCT based on [5]

While performing the main task, different kinds of errors (e.g., late or incorrect execution) can occur at various stages of cognitive processing. For instance, at the stage of perception, the operator may fail to direct visual attention to the relevant areas. Consequently, an approaching track may not be perceived. If the operator attends to the relevant area but does not process the stimulus or cannot discriminate the object from its environment, an error during the stage of object detection and discrimination occurs. At the decision-making stage, the incorrect application of decision rules can lead to decision errors. Finally, errors can also occur during the stage of task execution, e.g., by selecting the wrong response or omitting a response altogether.

The amount and complexity of the cognitive processes involved in tasks like the WCT require a performance mechanism that adequately represents the operator’s performance. However, the original performance assessment employed in the WCT neither considers the speed of task completion, nor the type and amount of errors of the operator. For example, in vigilance tasks with temporal components like reaction time, a single overall performance score cannot consider the speed-accuracy tradeoff that describes the process of sacrificing accuracy for speed in task execution [6]. Specifically, accuracy cannot be defined by a single numerical value in complex tasks like the WCT due to the various sources of possible errors and mistakes.

For our purposes of creating affect-adaptive interaction, these are serious limitations. Trading accuracy for faster task completion (speed-accuracy tradeoff) is unfavorable in a command-and-control (C2) task where errors can have fatal consequences. An adaptive system could employ strategies to shift the internal criterion of the operator to complete tasks with more accuracy.

Quantitative analyses of the different error types allow for an in-depth understanding of the decision-making processes of the operator. With a scoring approach that captures the multidimensionality of performance decrements, researchers can exploit the advantages of an objective, quantitative assessment method while learning more about the operators’ decisions that lead to the performance decrement.

In section 2, Signal Detection Theory (SDT) is introduced as theoretical background. Section 3 describes the scoring mechanism as it was developed for the chosen task environment followed by a discussion of lessoned learned and a generalized scoring system in section 4. Section 5 sums up the results, provides a conclusion and outlines future work.

II. SIGNAL DETECTION THEORY

Signal Detection Theory (SDT) provides a theoretical rationale on which quantitative measurement techniques in a wide range of application domains can be based on. Originating from signal detection in psychophysics [7], the theory successfully explains phenomena in the study of visual search [8], recognition memory [9], decision making in supervisory control [10], air combat training [11], essay grading [12] or social anxiety [13]. All these domains seem unrelated at first sight. However, the application of signal detection theory lead to a greater understanding by quantifying the underlying cognitive and behavioral processes.

SDT describes the process of detecting a signal [7] that can either be present or absent as well as detected or missed. This results in four possible outcomes, namely Hit, Miss, Correct Rejection or False Alarm (see Table I).

TABLE I. DECISION-MAKING IN SDT

		Signal	
		Present	***Absent***
Response	***Present***	Hit	False Alarm
	Absent	Miss	Correct Rejection

Based on that categorization, the operator’s criterion and ability to discriminate the signal (+ noise) from a noise-only condition can be determined. The operator’s tendency to exhibit a response independent from the presence of the nature of the signal is referred to as the criterion. An operator with a liberal criterion has higher False Alarm and Hit rates, whereas an operator with a conservative criterion has higher Correct Rejection and Miss rates. Discriminability is defined as the number of correct decisions (Correct Rejection and Hit) relative to the number of incorrect decisions (Miss and False Alarm) [14].

 ISBN: 978-1-61208-870-9

III. THE SCORING MECHANISM

We used signal detection theory as a basis in the development of a detailed scoring mechanism to assess the operator's response bias and performance across multiple performance criteria in a command-and-control environment (see section 3A). The first application of the scoring approach is described in section 3B.

A. Command-And-Control Tasks

The scoring mechanism was implemented in the Rich And Adaptable Test Environment (RATE), a modular and scalable task environment developed by Fraunhofer FKIE that allows for flexible design of experimental tasks. One instantiation of RATE is the described command-and-control task. This setup, coined RATE for C2, was used to investigate the relationship between performance and emotion in a command-and-control task [15]. In accordance with the WCT described in section 1A, the operator's task in RATE for C2 is to identify unknown tracks on a simulated radar screen and categorize them into neutral, hostile, or friendly tracks. Furthermore, hostile tracks that enter certain ranges around the own ship must be warned or engaged, respectively.

The operator's performance was measured by accuracy and speed in task completion. To assess performance at the subtask level, negative and positive scores for each category (accuracy and speed) were assigned to every subtask (identification. warning and engagement). The division into separate positive and negative scores is necessary to cover all categories of the SDT, as described in section 2.

The accuracy scores are based on the correctness of the operator's action. For instance, the engagement of a hostile track gains points on the positive accuracy score, whereas engagement of a neutral or friendly track increases the negative accuracy score. Similarly, positive and negative speed scores were used to capture the temporal performance aspects based on response time. For instance, timely engagement of a hostile track leads to an increase in the positive speed score, whereas a missed or delayed engagement of a hostile track increases the negative speed score. Figure 3 demonstrates how the subtasks of the WCT relate to the assigned performance scores.

The criterion of the operator is determined by the relationship between positive and negative scores of each category (accuracy and speed). An operator with a liberal criterion would tend to engage many tracks – including friendly or neutral ones - leading to an increase in positive as well as negative scores of accuracy. This corresponds to a high Hit and False Alarm rate in SDT. In contrast, a conservative operator would be hesitant in engaging tracks, leading to an increase in negative speed score. This corresponds to a high Miss Rate in SDT.

Insights into the cognitive processes associated with the subtasks of RATE for C2 arise from the analysis of individual scoring components. Problems with visual attention and perception, as well as object discrimination become evident in a high negative speed score for the subtasks of identification, warning, and engagement.

A high negative accuracy score results from problems occurring during the stage of decision-making and rule application. For instance, if the operator assigns a false identification to the corresponding track the negative accuracy score increases. This again results from incorrect interpretation of the IFF code.

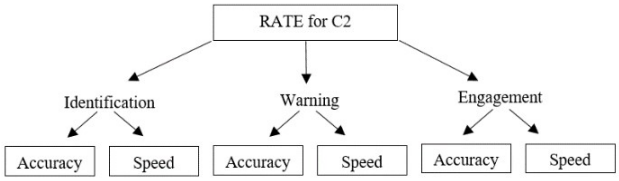

Figure 3. Subtasks of RATE for C2 and performance scores

Furthermore, a high negative speed score in combination with a high positive accuracy score indicates problems within the motor task execution.

To calculate total subtask and task performance, we deducted the negative from the positive score because high positive and low negative scores display high performance. However, achievable scores may vary greatly across conditions or scenarios. For example, in our task environment, difficulty levels were determined by the total number of tracks and the relative proportion of hostile tracks. In order to be able to compare absolute performance scores across conditions and scenarios it is necessary to normalize the absolute score. Our approach was to divide the absolute delta between positive and negative scores by the maximum achievable score within each category (see Table II), resulting in a comparable performance.

TABLE II. THE SCORING MECHANISM FOR ACCURACY AND SPEED ADAPTED TO A C2-TASK

Individual Scoring Components	Conditions
Accuracy	
Positive Score	Correct Decisions: Correct Identification, Correct Warning of Hostile Tracks, Correct Engagement of Hostile Tracks
Negative Score	Incorrect Decisions: False Identification, False Warning (friendly/neutral Track), False Engagement (friendly/neutral Track)
Total Score	Positive Score - Negative Score
Max Score	∑ Achievable Positive Scores
Performance	Total Score / Max Score
Speed	
Positive Score	Correct Decisions: Identification in <=30 seconds, Warning of Hostile Tracks in <=20 seconds, Engagement of Hostile Tracks in <=10 seconds
Negative Score	Incorrect Decisions: Identification in >30 seconds, Warning of Hostile Tracks in >20 seconds, Engagement of Hostile Tracks in >10 seconds

Individual Scoring Components	Conditions
Total Score	Positive Score - Negative Score
Max Score	∑ Achievable Positive Scores
Performance	Total Score / Max Score

All described scores were generated separately for each subtask (identification, warning, engagement). An overall total score is the sum of all subtask total scores. An operator's overall performance for the experiment is calculated along the lines of subtask performance (by dividing the experiment's total score by the maximum achievable points across all subtasks).

B. *Application of the Scoring Approach*

The scoring mechanism was used in a study with Fifty-one ($N = 51$) subjects aged 18 to 57 years ($M = 32.75$, $SD = 9.8$) to examine the relationship between performance and emotion in the command-and-control task [15] described above. Task load was modulated across scenarios by varying the total number of tracks and the relative proportion of enemy tracks. This approach was based on the cognitive task load model validated with a command-and-control task by de Greef and Arciszewski [16]. Cognitive task analysis and the review of similar tasks covered in literature helped us to identify and to rank all relevant subtasks, their priorities and dependencies, as well as all possible behaviors associated with each subtask. Figure 4 illustrates that our implementation of the described scoring mechanism was sensitive to task load, as the overall performance decreased with higher task load.

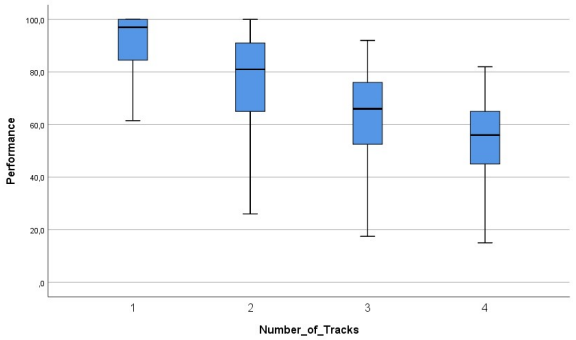

Figure 4. The performance score was sensitive to the number of tracks within a scenario (1=6 tracks, 2= 12 tracks, 3= 18 tracks, 4= 24 tracks)

IV. DISCUSSION

The developed scoring system offers several benefits for measuring human performance in complex task environments. Using "accuracy" and "speed" as performance criteria, we were able to gain insights into the cognitive processes associated with specific subtasks. The scoring approach itself does not isolate the specific information processing resource involved but it provides separate scores for each subtask and all performance criteria that can be mapped to cognitive processing steps. As mentioned above, problems with visual attention and perception as well as object discrimination were reflected in a high negative speed score whereas problems occurring during the stage of decision-making and rule application lead to a high negative accuracy score. Thus, careful choice of categories and/or tasks allows researchers to map subscores to cognitive processes and determine possible causes of observed behaviors.

However, a profound understanding of the experimental task and the cognitive processes involved is required to exploit these benefits. For example, being able to break down subtask performance and consider the hierarchical order of subtasks avoids misleading performance scores, but finding adequate ways for dealing with task omissions and conditional subtasks proved challenging.

The following lessons learned are meant to raise awareness for issues we encountered and provide possible ways to address them.

A. *Lessons learned*

1) Task structure: In order to define the actions leading to increases in positive and negative scores, respectively, decision trees have emerged to be a valuable tool to test the logical order of every possible subtask sequence and its associated scores.

2) Subtask priority: The priority of a subtask in the context of the overall task can be reflected in the amount of points earned on the corresponding positive/negative score. This allows the scoring mechanism to be adapted to other tasks environments, even beyond the command-and-control domain. Keeping in mind the research question and hypotheses of the study also helped to assess the relevance of subtasks and the relationship between them.

3) Conditional subtasks: Caution is advised when scoring conditional subtasks. Conditional subtasks are subtasks that occur in dependence of the outcome of a previous subtask. For example, in the described command-and-control task, only hostile tracks have to be engaged. If the operator incorrectly identifies a friendly track as hostile and then engages it (correctly, from the operator's point of view), no points should be awarded, otherwise he could earn more points than the maximum possible. Therefore, correct actions in conditional subtasks must not be rewarded if the action was only correct because of a preceding error. Whether incorrect actions lead to points on the negative score should be determined in the specific task context.

4) Task omissions: When determining point allocation, omission of necessary actions should neither lead to points on the positive nor on the negative score in order to separate omission errors from correct or incorrect explicit behavior. In case of the accuracy score, points on the positive score represent Hits in SDT and points on the negative score represent False Alarms in SDT as described above. Giving points for omissions would falsify this representation of SDT categories. This, however, does not apply if the omitted

 ISBN: 978-1-61208-870-9

action represents incorrect behavior. For example, if the performance criterion is speed, operators should be given points on the negative score when the omission represents a missed identification.

5) Normalization: Normalization of the absolute performance score enabled comparisons of operator performance across conditions or scenarios, and even across different experiments. With a normalized performance score, it is possible to capture and analyze changes in performance over time (e.g., to compare implemented usability improvements or new interaction mechanisms). The impact of changes or improvements can then be analyzed at the task and even at subtask level. Test-retest reliability is ensured because the calculation of the score is independent from any dynamic components except the actions of the operator himself.

B. Use With Other Tasks

The developed scoring approach is not limited to use in command-and-control environments. It could also be generalized and adapted to other tasks that consist of multiple subtasks, including hierarchical task structures. Validation is still pending but the generalized concept is described below.

As a first step, applicable performance criteria must be determined for the task at hand. For our task detailed above, we chose accuracy and speed but there may be other options. In our case, the operator has the option of performing a subtask correctly or incorrectly depending on the considered category. For instance, in the case of speed, correct means "within a time limit" and incorrect means "outside the time limit" but the scoring approach is not limited to these categories. The scheme provided in Table II can be adapted and enhanced to calculate performance metrics for each criterion and each subtask in other task environment.

To assess multiple performance criteria, the above procedure can be repeated for each criterion separately. Performance across all subtasks can be calculated by dividing the sum of all subtask total scores by the sum of all subtask max scores (see Table III). It is also possible to quantify overall performance across all subtasks and categories.

TABLE III. GENERALIZED SCORING MECHANISM FOR TWO CATEGORIES AND TWO SUBSTASKS

Scoring-Components	Conditions
Category A (i.e., accuracy)	
Subtask 1	
Positive $Score_{A1}$	Correct Decisions
Negative $Score_{A1}$	Incorrect Decisions
Total $Score_{A1}$	Positive $Score_{A1}$ - Negative $Score_{A1}$
Max $Score_{A1}$	∑ All possible achievable Positive $Scores_{A1}$
$Performance_{A1}$	Total $Score_{A1}$ / Max $Score_{A1}$
Subtask 2	
Positive $Score_{A2}$	Correct Decisions
Negative $Score_{A2}$	Incorrect Decisions
Total $Score_{A2}$	Positive $Score_{A2}$ - Negative $Score_{A2}$
Max $Score_{A2}$	∑ All possible achievable Positive $Scores_{A2}$
$Performance_{A2}$	Total $Score_{A2}$ / Max $Score_{A2}$
Overall subtasks in category A	
Total $Score_{A}$	Total $Score_{A1}$ + Total $Score_{A2}$
Max $Score_{A}$	Max $Score_{A1}$ + Max $Score_{A2}$
$Performance_{A}$	Total $Score_{A}$ / Max $Score_{A}$
Category B (i.e., speed)	
Subtask 1	
Positive $Score_{B1}$	Decisions made within 15 seconds.
Negative $Score_{B1}$	Decisions made in more than 15 seconds.
Total $Score_{B1}$	Positive $Score_{B1}$ - Negative $Score_{B1}$
Max $Score_{B1}$	∑ All possible achievable Positive $Scores_{B1}$
$Performance_{B1}$	Total $Score_{B1}$ / Max $Score_{B1}$
Subtask 2	
Positive $Score_{B2}$	Decisions made within 15 seconds.
Negative $Score_{B2}$	Decisions made in more than 15 seconds.
Total $Score_{B2}$	Positive $Score_{B1}$ - Negative $Score_{B1}$
Max $Score_{B2}$	∑ All possible achievable Positive $Scores_{B2}$
$Performance_{B2}$	Total $Score_{B2}$ / Max $Score_{B2}$
Overall subtasks in category B	
Total $Score_{B}$	Total $Score_{B1}$ + Total $Score_{B2}$
Max $Score_{B}$	Max $Score_{B1}$ + Max $Score_{B2}$
$Performance_{B}$	Total $Score_{B}$ / Max $Score_{B}$
Overall categories	
Total Score	Total $Score_{A}$ + Total $Score_{B}$
Max Score	Max $Score_{A}$ + Max $Score_{B}$
Performance	Total Score / Max Score

V. CONCLUSION

Measuring human performance in complex task environments is a challenge, especially when multiple subtasks of varying priority are present or when subtasks depend on one another, resulting in a hierarchical task structure. With the reported scoring mechanism, we addressed some of these challenges and illustrated an approach to quantitatively assess human performance in complex experimental tasks. We have begun and illustrated first steps to generalize the scoring mechanism developed for our task environment, using SDT as a theoretical foundation, so that it can be adapted to different task environments and applicable performance criteria.

One limitation of the reported approach is that the scoring mechanism is currently limited to subtasks with dichotomous responses (correct or incorrect). Whether (and how) more gradual responses could be mapped into the score will be investigated in the future.

ACKNOWLEDGMENT

The authors would like to thank Stephanie Hochgeschurz for valuable thoughts on the scoring mechanism and Mitja Galkin for his crucial contributions to the customization of the RATE task environment and the implementation of the scoring mechanism.

REFERENCES

[1] P. A. Ochieng, “An analysis of the strengths and limitation of qualitative and quantitative research paradigms”,. Problems of Education in the 21st Century, vol. 13, pp. 13-18, 2009.

[2] A. Queirós, D. Faria, and F. Almeida, “Strengths and limitations of qualitative and quantitative research methods”, European Journal of Education Studies, vol. 3, no. 9, pp 369-387, 2017.

[3] N. Ducheneaut, R. J. Moore, and E. Nickell, “Virtual “third places”: A case study of sociability in massively multiplayer games” Computer Supported Cooperative Work (CSCW), vol. 16, no. 1, pp. 129-166, 2007.

[4] Warship Commander 4.4; Computer Software; San Diego, CA: Pacific Science & Engineering Group; 2003.

[5] S. K. Card, T. P. Moran, and A. Newell, “The Model Human Processor: An Engineering Model of Human Performance” Handbook of Perception and Human Performance. vol. 2, Cognitive Processes and Performance, pp. 1-35, 1986.

[6] A. Osman et al., “Mechanisms of speed–accuracy tradeoff: evidence from covert motor processes”, Biological psychology, vol. 51, no. 2-3,pp. 173-199, 2000.

[7] D. M. Green and J. A. Swets, “Signal detection theory and psychophysics”, Wiley, 1966.

[8] P. Verghese, “Visual search and attention: A signal detection theory approach” Neuron,vol. 31, no. 4, pp. 523-535, 2001.

[9] J. T. Wixted, “Dual-process theory and signal-detection theory of recognition memory” Psychological review, vol. 114, no. 1, p. 152, 2007.

[10] A. Bisseret, “Application of signal detection theory to decision making in supervisory control The effect of the operator's experience”, Ergonomics, vol. 24, no. 2, pp. 81-94, 1981.

[11] J. L. Eubanks and P. R. Killeen “An application of signal detection theory to air combat training”; Human factors, vol. 25, no. 4, pp. 449-456, 1983.

[12] L. T. DeCarlo, “A model of rater behavior in essay grading based on signal detection theory” Journal of Educational Measurement, vol. 42, no. 1, pp. 53-76, 2005.

[13] L Yoon, J. W. Yang, S. C. Chong, and K. J. Oh, “Perceptual sensitivity and response bias in social anxiety: an application of signal detection theory” Cognitive therapy and research, vol. 38, no. 5, pp. 551-558, 2014.

[14] N. A. Macmillan, “Signal detection theory”, Stevens' handbook of experimental psychology, Wiley, 2002.

[15] A. Schmitz-Hübsch, S. M. Stasch, R. Becker, and S. Fuchs, “Personality Traits in the Relationship of Emotion and Performance in Command-and-Control Environments”, International Conference on Advances in Computer-Human Interactions. vol. 14, in press.

[16] T. D. Greef and H. Arciszewski, ”Triggering adaptive automation in naval command and control”, Frontiers in adaptive control, IntechOpen, 2009.

 ISBN: 978-1-61208-870-9