Eye-tracking parameters as a predictor of human performance in the detection of automation failures

Catrin Hasse, Dietrich Grasshoff, & Carmen Bruder
German Aerospace Center,
Aviation and Space Psychology, Hamburg,
Germany

Abstract

The increasing amount of automation in aviation systems requires that the operators monitor those systems appropriately. "Operators monitoring appropriately" (OMA) have been defined as those who monitor in a way that enables them to detect automation failures and resume control if automation fails. Identifying OMA reliably is a current objective for the selection of future aviation personnel. Eye-tracking data have been utilised to provide real-time measurements of visual and cognitive information processing. This raised the question of which eye-tracking parameters are important for differentiating between high performance and poor performance among operators. Previous studies had revealed time-sensitive eye-tracking parameters that help identify OMA who are prepared to resume control. This study dealt with finding eye-tracking parameters that help identify OMA who are able to detect automation failures. An experiment was conducted with 33 candidates for the DFS (Deutsche Flugsicherung GmbH). A simulation tool called "MonT" (Monitoring Test) was developed, which required test subjects to monitor an automatic process and register automation failures while eye movements were recorded. Results have revealed suitable eye tracking parameters that help differentiate between the participants' performance level in detecting failures. In the long term, MonT will be further developed with the aim of meeting the criteria for future selection tests.

Introduction

According to research on the future of aviation, such as the Single European Sky ATM Research Program (SESAR), operators will have to work with highly automated systems. Wickens, Mavor, Parasuraman and McGee (1998) concluded that automation might affect system performance due to the new skills that may be required, and that human operators might not have been adequately selected and trained to prepare for these changes.

In order to gather expectations about future tasks and roles, workshops were conducted with experienced pilots and air traffic controllers (Bruder, Jörn, & Eißfeldt, 2008). Findings from the workshop debriefings suggest that there is a

In D. de Waard, K. Brookhuis, F. Dehais, C. Weikert, S. Röttger, D. Manzey, S. Biede, F. Reuzeau, and P. Terrier (Eds.) (2012). Human Factors: a view from an integrative perspective. Proceedings HFES Europe Chapter Conference Toulouse. ISBN 978-0-945289-44-9. Available from http://hfes-europe.org

crucial new requirement for humans operating in man-machine settings: "operational monitoring." Operational monitoring includes using one's senses to follow meaningful information from various sources (e.g. an automated system) responsibly, even when there is no direct need for action. It involves being prepared to fully take over control of a system at any time, for example in the case of malfunction (Eißfeldt et al. 2009). Thus, the increase in automation requires operators monitoring appropriately (OMA). OMA are assumed to monitor in such a way as to enable them to detect system errors in time, and to take control if automation fails.

As the DLR's Department of Aviation and Space Psychology is responsible for the selection of pilots and air traffic controllers, one of its goals is to find criteria for identifying which candidates are suitable to become future operators.

Defining adequate monitoring performance: Devising a normative model

Niessen and Eyferth (2001) developed a model for an experienced air traffic controller's mental representation of a traffic situation. According to the model, adequate monitoring involves going through a monitoring cycle consisting of specific monitoring phases: orientation, anticipation, detection and recheck (ebd.). Based on this background, a theoretical model can be devised which describes the monitoring behaviour of OMA (figure 1). Thus, operators monitoring appropriately (OMA) are expected to orient themselves to automated system operations as well as to anticipate, detect and recheck them in time.

Recent research has tested the normative model with airline pilots and air traffic controllers. The monitoring behaviour of these experts supported the relations beween monitoring phases and performance data which were predicted by the model (Bruder, Grasshoff & Hasse, in press).

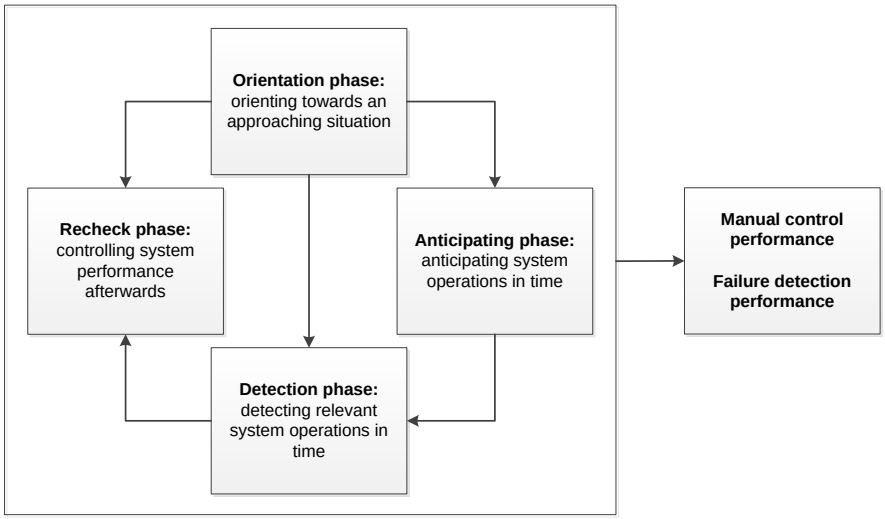

Figure 1: Normative model of phase specific monitoring behaviour

Measuring adequate monitoring performance: Eye tracking

Although the criteria for effective monitoring behaviour have been derived, until now there has not been a suitable way of measuring monitoring performance.

A variety of psychophysiological and imaging studies support the idea that eye movements offer an appropriate means for measuring the efficient and timely acquisition of visual information. For example, shifts in attention are usually reflected in the fixations (Findlay & Gilchrist, 2003). The fixation count can be used as a measure of a person's expectations and assumptions (Rötting, 2001). The fixation duration reflects the duration of information processing (Inhoff & Radach, 1998). Finally, the total gaze duration per AOI (area of interest) is a measure of the difficulty of recording the information viewed (Rötting, 2001).

When using eye tracking as a measurement of monitoring behaviour, this implies the following. First of all, OMA direct their gaze towards potentially relevant system operations at the right time, thus reflecting that they orientate themselves toward upcoming events. Furthermore, they anticipate the events directly before they happen, detect them when they happen, and recheck them afterwards. This raises the question of how OMA typically guide their eye movements. In particular, which scanning profile enables the operator to detect automation failures and assume control when necessary?

The aim is to identify suitable eye-tracking parameters which record the monitoring process and, at the same time, are related to the detection of automation failures.

Validating adequate monitoring performance: Performance measurements

Based on this line of thinking, eye tracking parameters that predict the ability to resume control were identified (Hasse, Bruder, Grasshoff & Eißfeldt, 2009b). Results indicate that the suitability of each parameter depends on the specific phase of the monitoring process. Gaze durations allow for differentiation between high and low performing subjects during orientation phases. In contrast, relative fixation counts are suitable for predicting monitoring performance during detection phases (Hasse, Grasshoff & Bruder, 2012).

So far, criteria for effective monitoring behaviour in relation to manual control have been identified, however, their relation to the ability to detect automation failures remains unclear. Thus, the question arises: which monitoring criteria are most important for identifying OMA who also have the ability to detect automation failures? In order to learn more about this relationship, the following hypotheses have been tested:

Hypothesis 1: While monitoring automated processes, adequate attention allocation during the orientation, anticipation, detection and recheck phases is related to the detection of automation failures.

Hypothesis 2: In terms of detecting automation failures, high performing operators differ from poorer performing operators in that they show adequate attention allocation during the orientation, anticipation, detection and recheck phases.

The hypotheses imply that the performance in detecting automation failures serves as a criterion for evaluating the quality of individual monitoring behaviour. Similarly, it is assumed that one's ability to monitor automation is indicative of one's performance in the detection of automation failures.

Method

An empirical study was undertaken with candidates for a professional training program at DFS (Deutsche Flugsicherung GmbH). Its purpose was to test the theoretical model of monitoring behaviour, i.e. its postulated monitoring phases and their relationships to the detection of automation failures. A simulation tool was developed providing both the assessment of monitoring performance and success at detecting automation failures. Diverse eye movement parameters were recorded to measure monitoring behaviour.

Simulation equipment/Simulation tool

A simulation tool called the “Monitoring Test” was developed to enable the assessment of monitoring behaviour and detection of automation failures. Since the tool is a simplified and abstract simulation of traffic flow, the test subjects need no prior experience as a pilot or air traffic controller. The traffic flow simulation can be controlled either automatically or manually by using input devices. The task of both the automated system and the human operator is to bring all current values into agreement with target values (for further information, see Hasse, Bruder, Grasshoff & Eißfeldt, 2009a). Objects (the arrows in figure 1) move at four second intervals.

24 scenarios were presented in the present study; in half of them, a single automation failure occurred (twelve malfunctioning scenarios). In the other twelve scenarios, the automatic system worked accurately (twelve distractor scenarios).

Figure 1: Air traffic flow display of the MonT simulation (screenshot).

Eye tracking equipment

Eye movements were recorded by the Eyegaze Analysis System, manufactured by L.C.T.. The system was combined with the simulation tool MonT to ensure that both systems use the same timestamp. The raw data was processed using NYAN software, developed by Interactive Minds. Subjects were seated in front of a 19-inch LCD computer display at a distance of approximately 60 cm.

Test subjects

The experiment was conducted with a sample of 33 applicants for air traffic control training at DFS (Deutsche Flugsicherung GmbH). They were 18 to 25 years old (M=19.48, SD=2.02), 61% were male. 60% claimed to have experience with strategy games. Experiments were conducted in conjunction with the regular selection process at the German Aerospace Center without influencing the selection outcome. Participants received 20 €.

Procedure

The experiment started with detailed instructions and four exercises. Participants were informed that they would be working on an automated traffic-flow simulation. The task was to monitor the automated system and detect false input devices in the system. Subjects were told to indicate the false input devices by locating them on the screen as soon as possible and clicking on them. After the briefing, participants had to monitor 24 traffic scenarios, each of which lasted two to three minutes. During the scenarios, the traffic moved dynamically. Every scenario began with an orientation phase, where the display was frozen. The orientation phase was meant to enable the participant to form a mental image of the simulation before the traffic

would start flowing dynamically. The duration of the orientation was variable, allowing the subject to take as long as necessary to orient himself. Finally, participants were asked about their impressions of the experiment.

Measurements

Eye tracking data and failure detection performance data were used as dependent variables.

Two groups of eye-tracking data were used to measure monitoring performance. The first group of data consisted of the total and relative fixation counts (rgd), meaning the ratio between the number of visual fixations on defined areas of interest (AOIs) and all fixations within a given time frame. Secondly, the total and relative relative gaze durations (rgd) on predefined AOIs were recorded.

AOIs directly represent system operations executed by the automatic system. Thus, perceiving these AOIs at the right time should indicate ideal monitoring behaviour. We distinguished between AOIs where the automation failure occurs (*relevant AOIs*), and AOIs representing events that could potentially encounter an automation failure (*all AOIs*). As orientation, anticipation, detection and recheck are only possible within certain time frames within a scenario, every scenario was divided into sections. Each time frame represents a monitoring phase and is characterised by AOIs that are necessary for monitoring adequately during this phase.

Regarding failure-detection performance, frequency and response time of failure detection were used, as well as the number of false alarms. Subjects indicated that they had detected a system failure by clicking on the input device where the failure occurred.

Additionally, the perceived effort and the individual duration of the orientation phase were recorded. After every scenario, perceived effort was measured by SEA scale (Eilers, Nachreiner, Hänecke & Schütte, 1986) from zero (no effort) to 220 (very high effort). During the orientation phase, when the display was frozen, participants could orient themselves for as long as they wanted. The duration of the orientation phase was analysed as an indicator of the difficulty to visualise the scenarios.

Results

All twelve malfunctioning scenarios were analysed. First analyses indicated that six of them differentiate sufficiently between participants. Therefore, further analysis focused on these six scenarios. The relationship between eye gaze data and performance data was examined. In addition, data groups with distinctly positive or negative results in failure-detection performance were identified and compared in terms of their to eye movement parameters.

Performance data

At 93% (SD=23), the failure detection rate was high for all 6 scenarios. However, the failure detection rate for individual the individual scenarios ranged from 50% to 100%, thus showing that some scenarios differered better between "detectors" and "non-detectors" than others. Automation failures were detected with a response rate of 2.43 seconds (SD=.15). False alarms happened in 26% of the scenarios (SD=3.0). On average, the subjective effort was rated at 44.35 (SD=4.62). On average, participants needed 16.27 seconds (SD=1.9) to orient themselves during the orientation phase.

Relationship between monitoring behaviour and failure detection performance

In order to examine the relationship between eye tracking data and performance data, scenarios were analysed both together and separately. In both instances, AOI-specific fixation counts and gaze durations during monitoring phases correlated with failure-detection parameters, such as failure detection frequency, false alarms and response time. Eye tracking parameters were analysed with respect to certain AOI groups (*relevant AOIs* and *potentially relevant AOIs*). This paper focuses specifically on eye tracking parameters that significantly correlate with failure-detection frequency.

Taking all 6 scenarios together, relative fixations counts (rfc) on *potentially relevant AOIs* during anticipation phases correlate significantly with failure detection frequency ($r=.39$; $p<.05$). The greater the proportion of fixations that fall on all potentially relevant AOIs during anticipation phases, the higher the failure-detection frequency. In addition, a significant negative correlation with relative gaze duration (rgd) on *potentially relevant AOIs* during anticipation phases was found ($r=.41$; $p<.05$). The longer the gaze remains on potentially relevant areas of interest during anticipation phases, the higher the frequency of failure detection. During detection phase, total fixations counts (tfc) on *relevant AOIs* correlated significantly with failure detection frequency ($r=.42$; $p<.05$). The greater the fixation count on relevant areas during detection phases, the higher the failure-detection frequency (figure 2).

In addition, orientation duration significantly correlates with the failure-detection frequency ($r=-.37$; $p<.05$). That is, the longer that subjects orient themselves towards upcoming events, the poorer they are at detecting automation failures afterwards.

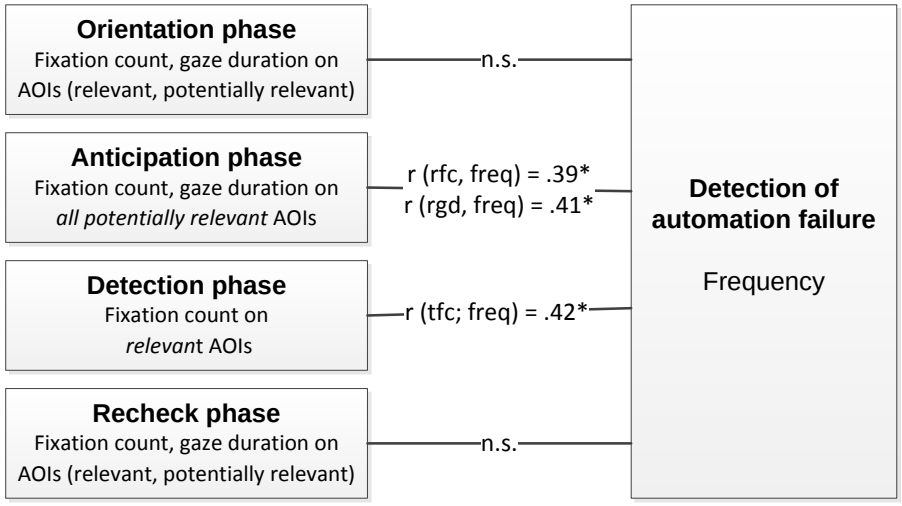

*Figure 2: Correlations between fixation count (rfc=relative fixation count, tfc=total fixation count) or gaze duration (rgd=relative gaze duration, tgd=total gaze duration) and failure detection frequency (n=33, * p<.05, ** p<.01).*

Group comparisons

In order to get a better understanding of the link between monitoring and manual performance, participants were divided into two groups according to their failure-detection frequency using a median split (high performers and low performers). Unpaired T-tests were used to compare groups with regard to several eye tracking parameters. The focus was on finding the eye tracking parameters that best account for individual differences in failure-detection frequency.

During anticipation phases, high performers demonstrate significantly higher fixation counts (rfc) on all potentially relevant AOIs than low performers do [$t(31)=2.22$; $p < .05$)]. In addition, high performers demonstrate significantly longer gaze durations (rgd, tgd) on all potentially relevant AOIs than low performers during the anticipation phase [$t(31)=2.30$; $p < .05$)] [$t(31)=2.65$; $p < .05$)]. That is, participants with a higher failure-detection rate gazed significantly more frequently and longer at all potentially relevant areas than poorly performing participants. During detection phases, high performers demonstrate significantly higher fixation counts (tfc) on relevant AOIs than low performers [$t(31)=2,34$; $p < .05$)]. That is, participants with a higher failure-detection rate looked significantly more frequently at relevant areas than poorly performing participants (figure 3).

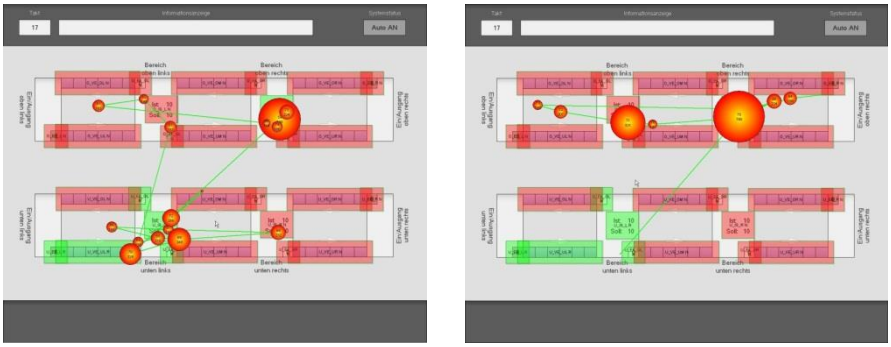

Figure 3: Distribution of fixations as scan paths during detection phase within a scenario. Left: Participant with good failure detection performance and adequate attention allocation: Fixations fallon relevant AOIs (light grey shades areas). Right: Participant with poor failure detection performance and inadequate attention allocation: Fixations miss relevant AOIs.

Since the orientation period varied between the candidates, fixation duration and frequency were confounded with the duration of the orientation period. Because of this, eye-tracking parameters were standardised according to orientation duration. The group comparison between subjects who performed well and those who performed poorly showed a significant effect on the standardized number of fixations on relevant AOIs [$t(31)=2.09$; $p < .05$)]. That is, during the orientation phase, participants with a high failure-detection rate show adequate attention allocation on relevant AOIs, whereas poorly performing participants failed to focus on upcoming relevant events.

Discussion

The present study focused on validating a theoretical model of adequate monitoring behaviour with candidates for selection of air traffic controller trainees. It was assumed that accuracy in monitoring, as defined by a theoretical model, is directly linked to one's competence in detecting automation failures.

The following conclusions can be made: Monitoring adequately (i.e. according to the model) enables operators to detect automation failures. However, some monitoring phases seem to be more important than others with respect to the ability to detect automation failures. Regarding the measurement of monitoring behaviour, appropriate eye tracking parameters were able to be identified, i.e. parameters that are associated with failure-detection performance. However, once again it depends on the monitoring phase which eye-tracking data best account for individual differences in failure-detection accuracy.

Operators monitoring in accordance with the normative model detect automation failures.

Generally, candidates with adequate monitoring behaviour detect automation failure more frequently. Thus, results are comparable to findings by Hasse, Bruder,

Grasshoff and Eißfeldt (2009a; 2009b) where manual control served as the criterion for evaluating the quality of monitoring, instead of failure detection.

However, some monitoring phases were more important for failure-detection performance than others. Participants who adequately distributed their attention during anticipation and detection phases showed better failure-detection performance than participants with random attention allocation. On the other hand, during recheck phases, this effect disappeared. Consequently, monitoring adequately during anticipation and detection phases seem to be more important for failure detection than monitoring adequately during recheck phases.

Similar to findings by Hasse, Bruder, Grasshoff and Eißfeldt (2009a; 2009b) the data from this study proved that the detection phase is highly important. Viewing the results of both studies, one can say that fixations on relevant AOIs in the phase where critical events take place lead to a higher probability of detecting failures and to a higher probability of resuming control successfully.

Contrary to the theory, both studies indicate that recheck seems to be less important than predicted by the theoretical model. In contrast to real air traffic control, no critical events happened during the recheck phases in the experiment. Candidates may have learned from this and reduced their monitoring during recheck phases.

In contrast to the former study (ebd.), attention allocation during the anticipation phase significantly correlated with failure detection, whereas it did not correlate with resuming control. An additional difference to the previous study was that attention allocation during the orientation phase had been shown to correlate significantly with resuming control, whereas in this study it did not correlate with failure detection. Thus, anticipating automatic events might be more important for the detection of automation failures than it is for the ability to successfully resume control when automation fails. This would mean that the criterion (failure detection vs. taking over manual control) influences which monitoring behaviour works well and which phase is particularly important.

However, this kind of conclusion could not be made for the orientation phase, since both experiments operated with different designs for the orientation phase. Whereas in this study participants had to decide how long they oriented themselves, the orientation phase in the former study was fixed. Giving every subject enough time to grasp upcoming events might have reduced the selectivity of the orientation phase in this experiment. In order to avoid this, further experiments should use a fixed duration for orientation.

It depends on the monitoring phase which eye-tracking data best account for individual differences in failure-detection accuracy.

Another result was that the phases influence which kind of AOI is important for failure detection. During *anticipation phases*, fixations that fall on *potentially relevant AOIs* are significantly linked to failure-detection frequency. In contrast, during *detection phases*, the gaze has to be directed towards *relevant AOIs* in order to correlate significanty with failure-detection performance. This could mean that in

order to anticipate automatic events and failures adequately, all potentially relevant areas should be taken into account. However, in order to detect automation failures at the time that they occur, clearly it is only important to focus on the relevant areas, i.e. where the failure occurs.

In summary, testing monitoring behaviour using dynamic simulations based on eye movements is an innovative approach that enables the development of new methods of personnel selection. We identified time sensitive eye-tracking parameters to serve as basis for identifying OMA in future selection processes. In this regard, we have shown that eye tracking parameters are predictive of failure-detection performance. Thus, the monitoring test (MonT) can be introduced as an effective tool for investigating human performance in future ATM scenarios.

Current research is focused on optimizing scenario difficulty. Scenarios with medium difficulty might generate stronger relations between eye tracking parameters and failure-detection performance. In addition, results from this sample are being compared to results from a sample of experts consisting of experienced pilots and air traffic controllers. That project investigates how air traffic controllers and pilots can be distinguished from job candidates in terms of their ability to monitor dynamic traffic situations. Further research will focus on team monitoring behaviour by assessing the monitoring and failure-detection behaviour of two participants who monitor traffic situations together. In this context, a team version of MonT will be developed in order to enable the assessment of team monitoring performance.

References

Bruder, C., Grasshoff, D. & Hasse, C. (in press). A model for the future: Operators monitoring appropriately. In *Proceedings of the 30th Conference of the European Association for Aviation Psychology* (pp 24-28). September 2012 in Villasimius (Sardinia), Italy.

Bruder, C., Jörn, L. & Eißfeldt, H. (2008). Aviator 2030 - When pilots and air traffic controllers discuss their future. In A. Droog, and T. D'Oliveira (Eds.), *Proceedings of the 28th Conference of the European Association for Aviation Psychology* (pp. 354-358), Valencia, Spain: EAAP.

Eißfeldt, H., Grasshoff, D., Hasse, C., Hoermann, H.-J., Schulze Kissing, D., Stern, C., Wenzel, J. & Zierke, O. (2009). *Aviator 2010 – Ability requirements in future ATM systems II: Simulations and experiments*. DLR Forschungsbericht 2009-28. Köln: DLR.

Findlay, J. & Gilchrist, I. (2003). *Active Vision.* Oxford University Press.

Eilers, K., Nachreiner, F., Hänecke, K. M. & Schütte, M. (1986). Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung. *Zeitschrift für Arbeitswissen-schaft, 40*, 215-224.

Hasse, C., Grasshoff, D. & Bruder, C. (2012). How to measure monitoring performance of pilots and air traffic controllers. *Proceedings of the symposium: Eye Tracking Research and Applications 2012*, Santa Barbara, USA.

Hasse, C., Bruder, C., Grasshoff, D. & Eißfeldt, H. (2009a). Future ability requirements for human operators in aviation. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics*, 8th International Conference, EPCE 2009, Held as Part of HCI International 2009, San Diego, USA (pp. 537-546). Berlin, Heidelberg: Springer.

Hasse, C., Bruder, C., Grasshoff, D. & Eißfeldt, H. (2009b). Future ability requirements for operators in aviation regarding monitoring. In A. Lichtenstein, C. Stößel and C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme*, 8. BWMMS, Fortschritt-Berichte VDI, Reihe 22, Nr. 29 (pp. 159-160). Berlin: ZMMS, TU Berlin.

Inhoff, A. W. and Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. In G. Underwood (Ed.), *Eye guidance in reading, driving and scene perception*, (pp. 29-53). München, Berlin: Elsevier.

Niessen, C. & Eyferth, K. (2001). A model of the air traffic controllers' picture. *Safety Science, 73,* 187-202.

Rötting, M. (2001). *Parametersystematik der Augen- und Blickbewegungen für arbeitswissenschaftliche Untersuchungen.* Aachen: Shaker.

Wickens, C. D., Mavor, A. S., Parasuraman, R. & McGee, J. P. (1998). *The future of air traffic control: Human operators and automation.* Washington, DC: National Academic Press.