Operators' Adaption to Unreliability of Alarm Systems: A Performance and Eye-Tracking Analysis

Linda Onnasch, Stefan Ruff, Dietrich Manzey
Berlin Institute of Technology
Berlin, Germany

Operators in complex environments are supported by alarm-systems that indicate when to shift attention to certain tasks. As alarms are not perfectly reliable, operators have to select appropriate strategies of attention allocation in order to compensate for unreliability and maintain overall performance. This study investigates how humans adapt to differing alarm-reliabilities. Within a multi-tasking flight simulation, participants were randomly assigned to four alarm-reliability conditions (68.75%, 75%, 87.5%, 93.75%), and a manual control group. In experimental conditions, one out of three subtasks was supported by an alarm-system. Compared to manual control, all experimental groups benefited from alarms in the supported task, with best results for the highest reliability condition. However, analyses of performance and eye-tracking data revealed that the benefit of the lowest reliability group was associated with an increased attentional effort, a more demanding attention allocation strategy, and a declined relative performance in a non-supported task. Results are discussed in the context of recent research.

INTRODUCTION

Alarm systems are a widespread technology in complex work environments used to support complex supervisory control tasks of operators. This is enabled by the attention-grabbing properties of alarm systems so that operators can be relieved from a continuous monitoring while staying in the loop as alerts inform them when to shift attention to a critical task (Pritchett, 2001). Benefits of such alarm systems can be described in terms of reduced workload and a performance increase in the alarm supported task as well as in concurrent tasks as operators gain more spare capacities, which can be reallocated (e.g. Bustamante, Anderson & Bliss, 2004).

However, the proposed benefits of this kind of automation can be off-set when alarm systems do not function properly. There are two different errors that can occur and have to be differentiated. The system can fail to alert the operator by missing critical events. On the other hand, the system may alert an operator too often as not every alert corresponds to a critical event. In this case the alarm system would produce false alarms (Swets, 1964; Green & Swets, 1966). Given these possible failures, operators' responses to alarms always imply a decision under uncertainty that is mainly based on their assessment how much they can rely on the alarm function.

According to Lee & See (2004) the most important perceivable characteristic for the calibration of reliance on automation (like alarm systems) is the system's reliability. I.e., the higher the alarm system's reliability, the more the operator can rely on the alarm and the less he is required to monitor the underlying data himself. In contrast, when reliability is low, the operator should more frequently monitor the underlying data, which is monitored by the alarm system, in order to compensate for the system's imperfection, particularly if the alarm tends to miss critical states.

How operators adapt their own monitoring behavior in case of available alarm systems or other decision support has been addressed in several studies (e.g. Parasuraman, Molloy & Singh, 1993; Wickens & Dixon, 2007). However, the results are mixed. For example, Bailey and Scerbo (2007) examined operators' adaption to a highly reliable support system that automatically indicated and resolved critical system states within a multi-task environment. Results indicate that monitoring of the supported task inappropriately decreased as a function of increasing system reliability. These findings support earlier results by Molloy and Parasuraman (1996) who also reported degraded monitoring efficiency in interaction with a highly reliable system. On the contrary, other studies support the assumption that operators are very well capable to adapt to changing reliability levels as well as to changes in initial levels of reliability, suggesting nearly optimal adaption strategies (e.g. Parasuraman et al., 1993, Wiegmann, Rich, & Zhang, 2001)

Though, in most of these studies the evaluation of monitoring performance was solely based on operators' performance alone (Parasuraman et al., 1993; Wiegmann et al. 2001; Bailey & Scerbo, 2007). This does not seem to be appropriate as the concept of an automated assistance or alarm system is to support the operator and to resume parts of the task; i.e. the task is performed jointly. Therefore, the joint human-automation performance should always be considered in order to evaluate performance consequences of operator behavior.

In accordance with this approach, Wickens and Dixon (2007) conducted a meta-analysis consisting of 22 studies with varying reliabilities. In contrast to most of the aforementioned research, they found a positive linear relation between the automation's reliability and the joint human-automation performance. However, below a reliability of 70% this compensation was associated with a disproportional effort, and performance got even worse than when working with no automation at all. Thus, compensation for unreliability seems to be possible only to a certain level.

Summarizing the scope of this research, it is still in question how operators can adapt their monitoring strategies to specific system characteristics. For high levels of reliability there is some support that people are not able to adapt properly resulting in inadequately low levels of own system monitoring. Additionally, the meta-analysis of Wickens and Dixon (2007) suggests that adaption to low reliabilities is challenging

 DOI 10.1177/1071181312561059

as well and often resulting in impaired overall performance. Based on this pattern of results, the first goal of the current study was to gain further insights into possible adaption strategies to alarm systems with respect to different levels of alarm reliability.

The second goal was to examine how this adaption proceeds in detail. For example, Lee & See (2004) propose a monotonous linear relation between automation capabilities and operators' trust in and reliance on the automation they are working with. Within their framework a diagonal line represents this relation where the level of trust matches automation capabilities. Everything above and below this diagonal describes mismatches of reliance and system characteristics. These assumptions are illustrated in Figure 1 (line a) with reliance as the behavioral realization of trust and reliability representing automation capability.

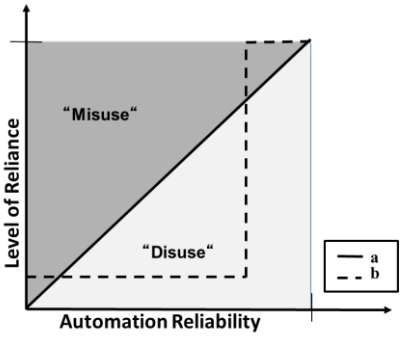

Figure 1. Relationship among operator's reliance and automation's reliability according to Lee & See (2004); Line a represents a linear relation, line b a dichotomization of behavior in dependence of reliability

However, empirical evidence still has to be provided. In most of the aforementioned studies, only relatively extreme levels of reliability were compared which do not allow for a detailed analysis of adaptive behavior. Therefore, the kind of relation between reliability and reliance still has to be clarified. There is some theoretical support for a linear relationship between reliability and reliance (Lee & See, 2004) but other shapes of relationship are also possible. For example, operators could tend to dichotomize their behavior in response to automation's reliability (line b). This dichotomization could explain operators' insensitivity to varying levels of reliability (Parasuraman et al., 1993) and the findings of inappropriately high or low reliance as response to alarm systems with high and low reliability, respectively. In this case, operators would just decide whether to rely or not on the automation.

METHOD

Participants

A total of 65 engineering students (18 female, 47 male) ranging in age from 19 to 32 (M = 23.6, SD = 2.3) participated in partial fulfillment of course requirements. None of the participants had any prior experience with the flight simulation task used in the study.

Apparatus: Microworld – MATB

The most recent version of the Multi-Attribute Task Battery (MATB, Miller, 2010) was used for the experiment. Compared to the original one developed by Comstock and Arnegard (1992), this version differs with respect to the programming environment which was changed from QBasic to MatLab and could therefore be used on Windows XP operating systems. All main functionalities, e.g. the user interface, remained unchanged. The MATB is a multitask flight simulation consisting of three concurrent but equally weighted tasks: A compensatory tracking task, a resource management task, and a system monitoring task.

In the compensatory tracking task, participants are required to keep a randomly moving cursor in the center target position by applying appropriate control inputs via joystick.

In the resource management task, participants have to compensate for fuel depletion by pumping fuel from four supply-tanks into two main tanks.

The system monitoring task consists of four engine gauges that participants have to monitor for randomly occurring abnormal values. These deviations represent system malfunctions, which have to be detected and reset by a corresponding key press. If a malfunction is not detected within 10 seconds the gauge resets automatically and the event is defined as a miss. In the current experiment, in every 10-minute period 16 malfunctions occurred, which had to be detected. Dependent on the experimental condition, this latter task was supported by a binary alarm system of varying reliability. When working with the alarm system a visual red alert appeared whenever a parameter deviated from the optimal level. Nevertheless, the identification of the affected gauge and the corresponding reset of the parameter still had to be done manually by the operator. According to the stages and levels taxonomy of automation proposed by Parasuraman, Sheridan and Wickens (2000), the alarm system was classified as a stage 2 automation (information acquisition and analysis) leaving action selection and action implementation within operators' responsibility. To determine possible automation benefits or, following Wickens and Dixon (2007), possible automation drawbacks with low reliability, a manual control group was additionally implemented. In this condition all three tasks had to be done manually, i.e. the monitoring task was not supported by the binary alarm system.

Design

The study used a two factorial design. The first factor (Reliability) was defined as a between-subject factor and consisted of four experimental groups and one manual control group. As a function of condition, the alarm reliability was set to 68.75%, 75.00%, 87.5% or 93.75%, respectively, by varying the number of critical signals that were missed by the system. That is, e.g. for the lowest reliability condition, 5 out of 16 malfunctions were not signaled by the alarm system and therefore had to be detected by the participants.

The second factor (Block) was defined as within-subjects factor. Every participant had to work with the MATB for three blocks in her / his condition. Every block lasted 10 minutes. A

total of 16 critical events occurred in the monitoring task during each block.

Dependent measures

Three different categories of dependent measures were analyzed: Eye-tracking, performance, and subjective data.

(1) Eye-tracking measures: To assess the impact of an alarm system's reliability on participants' reliance, the attention allocation strategies of participants were analyzed using eye-tracking measurements. This operational definition complies with Moray and Inagaki's (2000) assertion to evaluate operators' performance not only by fault detection but first and foremost by an analysis of their monitoring strategies.

Before the experiment started, three different areas of interest (AOI) were defined corresponding to the three different tasks participants had to perform: Compensatory tracking, resource management and system monitoring. For each AOI two variables were evaluated: The *relative fixation time* was defined as the time participants fixated an AOI relative to the overall fixation time of the three different AOIs. The *relative fixation count* captured all fixations within one AOI in relation to the fixations on the three predefined AOIs.

(2) Performance measures were defined according to the three tasks participants had to work on. For the system monitoring task the *percentage of detected alarm failures – human alone* was evaluated. Only groups with alarm support were taken into account for this measure. The *percentage of detected system failures – human + alarm system* was defined as the overall performance of the cooperative team, human and alarm system in detecting deviations on one of the four gauges.

For the tracking task as well as the resource management task the *root mean squared errors* (RMSE) were calculated. The RMSE for the tracking task was defined as the deviation from the central target position. The RMSE for the resource management task was defined in relation to an optimal tank level that had to be maintained in both main tanks.

(3) Subjective measures were evaluated for the *perceived reliability* of the alarm system, and the *subjective workload* ratings. The perceived reliability was evaluated by asking participants "How reliable was the system you worked with?". Responses had to be provided on a scale ranging from 0% to 100%. For the workload assessment the NASA-Task Load Index (Hart & Staveland, 1988) was used.

Procedure

Following an instruction on the MATB and a first calibration of the eye tracker, participants were familiarized with the three different tasks in a 10-minute practice session. Afterwards, they were randomly assigned to one of the five conditions and according to condition introduced to the alarm system when in an experimental group. Then the experiment started consisting of three 10-minute blocks. The NASA TLX followed every block; the perceived reliability rating was presented after the second block.

RESULTS

Perceived Reliability

In order to assess to what extent participants were able to correctly recognize the reliability of the alarm system they had to work with *perceived reliability* ratings were compared with the actual reliability of the alarm system using t-tests. α was adapted to a 20% level as no differences between perceived and actual reliability were expected (null-hypothesis testing).

For the 68.75% and 75% reliability condition there were no differences between actual and perceived reliability ($M_{68.75\%}$ = 66.77, $t(12)$ = -.48, p = .63; $M_{75\%}$ = 72.38, $t(12)$ = -.52, p = .61). However, participants in the two highest reliability condition systematically underestimated the actual reliability, $M_{87.5\%}$ = 80.08, $t(12)$ = -3.29, p < .007; $M_{93.75\%}$ = 87.08, $t(12)$ = -3.09, p < .01.

The accurate perception of the system's reliability presents an important precondition for any adaptive behavior, as the actual level of reliability first has to be recognized before people can adapt to it. The results suggest that this precondition was at least partially fulfilled.

Eye-Tracking

For the *monitoring task*, participants in the two highest groups (93.75% & 87.5%) showed a relatively low and stable *fixation time* across blocks. Albeit on a somewhat higher level, the 75% reliability group revealed a similar pattern in their fixation time on the monitoring task. In contrast to this, the manual group and the 68.75% condition had a very similar increase of fixation time through blocks. These effects are illustrated in Figure 2. Analyzed by a 5 (Reliability) x 3 (Block) ANOVA these findings were statistically supported by a significant Block effect ($F(1.78, 107.13) = 20.63$, p < .001), moderated by a Block x Reliability effect, $F(7.14, 107.13)$ = 2.46, p < .03.

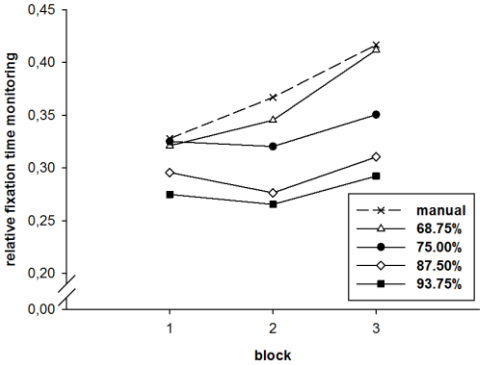

Figure 2. Effect of alarm reliability on the relative fixation time, AOI Monitoring

These results were mirrored in the *relative fixation time* for the *tracking task*. Inversely, the 93.75% and the 87.5% reliability groups had the longest fixations which only marginally changed over time whereas for the other groups a steep decrease was found which was most substantial for the 68.75% reliability condition (see Figure 3). This pattern was

statistically shown in significant main effects of Reliability ($F(4, 60) = 2.64$, $p < .05$) and Block, $F(1.68, 101.29) = 9.81$, $p < .001$, moderated by a Block x Reliability effect, $F(6.75, 101.29) = 3.62$, $p < .003$.

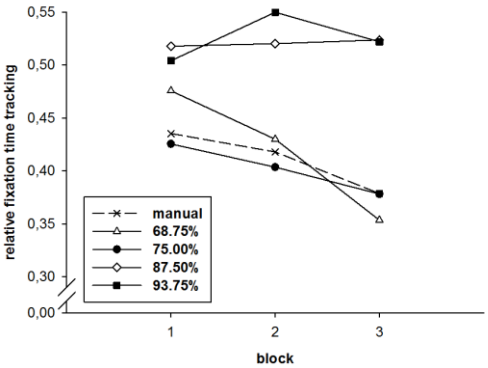

Figure 3. Effect of alarm reliability on the relative fixation time, AOI Tracking

Considering the *resource management task* the *relative fixation time* increased over blocks for the two lowest reliability groups (means 68.75%: 0.20, 0.22, 0.23; means 75%: 0.24, 0.27, 0.27) whereas a reverse effect was observed for all other conditions (means Manual: 0.23, 0.21, 0.20; means 87.5%: 0.18, 0.20, 0.16; means 93.75%: 0.22, 0.18, 0.18). The 5 (Reliability) x 3 (Block) ANOVA revealed a significant Block x Reliability effect, $F(7.16, 107.39) = 2.14$, $p < .05$. However, the resource management task was overall considerably less monitored than the other AOIs.

As a second variable *the relative fixation frequency* was assessed. For the *monitoring task* fixations increased independent of condition with time-on-task (M_{block1}= 0.30, M_{block2}= 0.30, M_{block3} = 34). Analyzed by a 5 (Reliability) x 3 (Block) ANOVA only the Block effect became significant, $F(1.75, 105.35) = 16.07$, $p < .001$.
Regarding the *tracking*, the 93.75% reliability group fixated this task most frequently, followed by the 87.5% condition (see Figure 4). The other groups fixated this AOI considerably less, revealing very similar results.

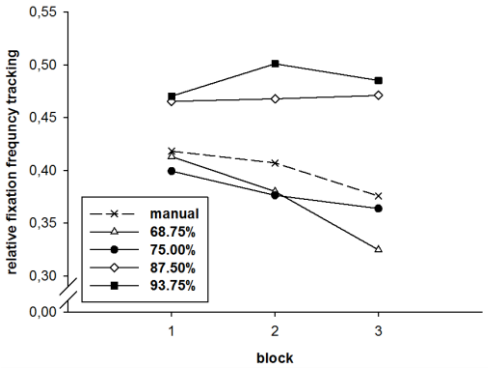

Figure 4. Effect of alarm reliability on the relative fixation count, AOI Tracking

Additionally, the overall fixation frequency declined over time with the 68.75% reliability group showing the steepest decrease. According to these findings, the 5 (Reliability) x 3 (Block) ANOVA revealed significant main effects of Reliability, $F(4, 60) = 2.69$, $p < .04$, and Block, $F(1.78, 107.11) = 7.63$, $p < .002$, which were both moderated by a Block x Reliability effect, $F(7.14, 107.11) = 3.08$, $p < .006$.
No effects were found for the *resource management task*.

Performance Measures

The *percentage of detected alarm failures* by participants in conditions with alarm support decreased as a function of reliability with lowest detection rates in the 93.75% condition ($M_{86.75\%}$ = 74.87%, $M_{75\%}$ = 70.51%; $M_{87.5\%}$ = 57.69%, $M_{93.75\%}$ = 51.28%). However, a 4 (Reliability) x 3 (Block) ANOVA did not reveal any statistical differences between the experimental conditions.

For the *percentage of detected system failures – human + alarm* all groups showed a time-on-task effect with better performance in the later blocks. Additionally, there was a clear alarm support advantage in detected system failures by human and automation compared to the manual control group (see Figure 5). A 5 (Reliability) x 3 (Block) ANOVA of these effects revealed significant main effects of Block, $F(2, 120) = 7.67$, $p < .002$, and Reliability, $F(4, 60) = 10.36$, $p < .001$.

Moreover, participants in the alarm supported groups adapted to the alarm system characteristics over time. No performance differences between these groups were observed in block 3 anymore. This was statistically supported by a significant Reliability x Block interaction, $F(8, 120) = 2.37$, $p < .03$). Additionally, a separate ANOVA comparing performance of the alarm supported groups for block #3 only revealed no significant differences ($p = .364$).

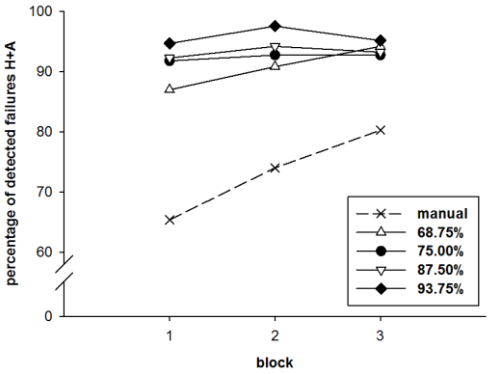

Figure 5. Effect of alarm reliability on detected system failures - human + alarm system

In the *tracking task* participants in the 68.75% reliability group started at a very high performance level revealing smaller RMSE ($M_{68.75\%}$ = 117.58). In contrast, the other groups showed slightly different but worse performance in the first block (M_{manual} = 131.78, $M_{75\%}$ = 136.05, $M_{87.5\%}$ = 144.57, $M_{93.75\%}$ = 137.76). However, whereas these groups could increase their performance with prolonged time, participants in the 68.75% reliability condition could not maintain their superior performance. This led to comparable performance levels in block #3 (M_{manual} = 124.62, $M_{68.75\%}$ = 126.94, $M_{75\%}$ =

126.78, $M_{87.5\%}$ = 127.58, $M_{93.75\%}$ = 129.55). A 5 (Reliability) x 3 (Block) ANOVA revealed a significant Block effect ($F(2, 120) = 7.84$, $p < .002$) that was moderated by a significant Block x Reliability interaction, $F(8, 120) = 3.59$, $p < .002$.

The RMSE analysis for the *resource management task* only showed a training effect as with increasing time-on-task all participants achieved better results represented in smaller RMSE (M_{block1} = 221.26, M_{block2} = 204.54, M_{block3} = 194.74). This was statistically supported by a significant Block effect, $F(1.2, 75.69) = 5.02$, $p < .03$.

Subjective Workload

Analysis of repeated subjective workload measures revealed a significant Block effect, $F(1.57, 94.27) = 8.96$, $p < .002$ that showed a reduction of workload over the three blocks (means: 53.51, 50.48, 47.97).

DISCUSSION

The main objective of this study was to investigate how capable human operators are in adapting strategies of attention allocation and multi-task performance to different reliability levels of alarm systems.

Two main conclusions can be drawn: (1) Results showed that below a critical alarm reliability level between 68.78% and 75% the maintenance of performance in the alarm supported task was associated with a disproportional attentional effort. This was revealed by the eye-tracking data. In contrast to the other alarm supported conditions, participants out of the 68.75% reliability group did not benefit from the alarm system and allocated as much attention to the supported task as the manual control group. This effect partially supports findings by Wickens and Dixon (2007) who propose a critical reliability cut-off around 70% below which automation support cannot be considered as helpful anymore. In the current study, performance did not decline in the supported task; but this could only be achieved by an increased cognitive effort and an attentional shift away from one of the concurrent tasks. This reallocation was accompanied by a relative performance decline in the related task. With respect to the fact that participants in the current study only had to work for 30 minutes with the system, the observed effects could be even more severe with a prolonged time-on-task. The additional attentional effort users had to invest may be hard to maintain. Ultimately, in terms of cognitive exhaustion, this overexertion could even lead to a complete performance breakdown (Hockey, 1997). Therefore, more research, especially longitudinal studies, is needed.

(2) Moreover, results reveal that for alarm reliabilities above the proposed critical level participants were sensitive to different reliability levels and adapted their own monitoring behavior in a monotonous manner to the alarm system‘s capability. Specifically, participants in the high reliability conditions (93.75% & 87.5%) monitored the alarm-supported task significantly less than participants in the two lower reliability conditions and used the gained cognitive resources for the concurrent tasks.

However, for the high reliability conditions this appropriate adaption was not mirrored in the perceived reliability ratings as participants systematically underestimated the alarm’s true reliability. This finding is in line with previous research (Wiegmann et al., 2001; Wiegmann & Cristina, 2000) and underlines the need to distinguish between subjective and behavioral performance data.

REFERENCES

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues of Ergonomics Science, 8*(4), 321-348.

Bustamante, E. A., Anderson, B. L., & Bliss, J. P. (2004) Effects of Varying the Threshold of Alarm Systems and Task Complexity on Human Performance and perceived workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*, 1948 – 1952.

Comstock, J. R, & Arnegard, R. J. (1992) *The multi-attribute task battery for human operator workload and strategic behavior research* (Technical memorandum No. 104174). Hampton, VA: NASA Langley Research Center.

Green, D. M., & Swets, J. A. (1966) *Signal detection theory and psychophysics*. New York: Wiley.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (139-183). Amsterdam, The Netherlands: Elsevier.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload. *Biological Psychology, 45*, 73–93.

Lee, J. D., & See, K. A. (2004) Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80.

Miller, W. D. (2010) The U.S. air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior (Technical report No. AFRL-RH-WP-TR-2010-0133). Wright-Patterson, OH: Air Force Research Lab. Retrieved from http://dodreports.com/pdf/ada537547.pdf.

Molloy, R. & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors, 38*, 311-322.

Moray, N. & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomic Science, 1*(4), 354-365.

Parasuraman, R., Molloy, R. & Singh, I.L. (1993). Performance consequences of automation induced “complacency”. *The International Journal of Aviation Psychology, 2*, 1-23.

Parasuraman, R., Sheridan, T. B. & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*(3), 286-297.

Pritchett, A. (2001) Reviewing the role of cockpit alerting systems. *Human Factors and Aerospace Safety, 1*(1), 5-38.

Swets, J. A. (1964). *Signal detection and recognition by human observers.* New York: John Wiley & Sons.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201-212.

Wiegmann, D. A., & Cristina, F. J.(2000) Effects of feedback lag variability on the choice of an automated diagnostic aid: a preliminary predictive model. *Theoretical Issues in Ergonomic Science, 1*, 139-156.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001) Automated diagnostic aids: The effects of aid reliability on users’ trust and reliance. *Theoretical Issues in Ergonomic Science, 2*(4), 352-367.