Benefits and Costs of Automation Support:
The Role of Function Allocation and Automation Reliability

vorgelegt von
Dipl.-Psych.
Linda Onnasch
geb. in Unna

von der Fakultät V - Verkehrs- und Maschinensysteme
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss

Vorsitzender: Prof. Dr. Manfred Thüring
Gutachter: Prof. Dr. Dietrich Manzey
Gutachter: Prof. Dr. Jürgen Sauer

Tag der wissenschaftlichen Aussprache: 13. Februar 2015

Berlin 2015
D83

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Des Weiteren versichere ich, dass die Darstellung des Eigenanteils an den in Co-Autorenschaft entstandenen Manuskripten der Wahrheit entspricht.

Berlin, den

Linda Onnasch

Acknowledgements

I would like to thank everybody who accompanied and supported me during my time at the chair of work, engineering and organizational psychology at the TU Berlin.

Special thanks go to my doctoral supervisor Prof. Dietrich Manzey for his scientific guidance, his patience, his excellent support and his personal effort in enabling this work. I have learned very much from him and this thesis is a result of this great mentoring. Furthermore, I would like to thank Prof. Jürgen Sauer for acting as second supervisor of my thesis.

I would like to express my gratitude to Prof. Chris Wickens with whom I worked on the meta-analysis publication. Chris' tremendous knowledge and experience in the field of human factors and automation research were invaluable. And even though the publication process was sometimes tough, the inspiring and also humorous conversations with him always encouraged me to go on.

Moreover, I would like to thank Dr. Rebecca Wiczorek for many inspiring discussions as well as the cooperation in our small science-projects that definitely broadened our understanding of many human factors concepts.

I want to thank Marcus Bleil for his great work regarding the technical implementation of the experimental microworld, the set-up of the eye-tracking equipment as well as his support in data aggregation. Furthermore, I would like to thank him for his motivational support and the interesting discussions during lunch breaks.

I would like to mention Nikolaus Rötting who particularly helped me with the eye-tracking set-up and Stefan Ruff for his support in collecting the data in the second experiment.

Furthermore, I want to thank my parents for their motivation and perpetual support. Last but not least, I would like to express my gratitude to Carsten Rasch for his scientific advice, his proof-reading of the thesis, his patience and encouragement at all times.

Zusammenfassung

Die vorliegende Dissertation hatte zum Ziel, das Wissen über den Einfluss von Automation auf die Gesamtsystemleistung als auch die kognitiven Anforderungen des menschlichen Operateurs weiter zu vertiefen. Die bisherige Forschung hat diesbezüglich zwei Automationscharakteristika hervorgebracht, welche maßgeblich bestimmen, ob die Unterstützung durch Automation einen positiven oder aber negativen Einfluss auf die Gesamtleistung darstellt: die Funktionsallokation (FA) zwischen Mensch und Automation und die Reliabilität der Automation. Mit Bezug auf diese zwei Aspekte haben Parasuraman, Sheridan und Wickens (2000) ein *flow chart* Modell publiziert, welches Entwicklern von automatisierten Systemen bei einer angemessenen Wahl der FA helfen soll. Dem Modell entsprechend sollte ein anfänglicher Automationsvorschlag durch sogenannte primäre Kriterien evaluiert werden. Diese umfassen mögliche Konsequenzen einer bestimmten FA auf der Leistungsebene, als auch in Bezug zu kognitiven Anforderungen an den Operateur. Sollte die initiale FA zwischen Mensch und Automation dieser Prüfung standgehalten haben, werden, dem *flow chart* Modell entsprechend, sekundäre Evaluationskriterien angelegt. Eine zentrale Rolle spielt hierbei die Reliabilität der Automation.

In Anlehnung an die konsekutive Struktur des *flow chart* Modells wurden im Rahmen der Dissertation drei Studien durchgeführt. Die erste Studie stellt eine Meta-Analyse dar (Studie I), welche den Einfluss von FA auf primäre Evaluationskriterien, Operateursleistung und kognitiven Aufwand, untersuchte. Die Ergebnisse zeigen, dass unter einwandfreier Funktionsweise der Automation mit steigendem Automationsgrad auch die Vorteile dieser Unterstützung in Bezug auf Leistung und operateursseitige Beanspruchung zunehmen. Falls es jedoch zu einem Automationsausfall kommt, steigt das Risiko negativer Konsequenzen in Bezug zur manuellen Aufgabenübernahme sowie dem Situationsbewusstsein des Operateurs mit steigendem Automationsgrad. Negative Konsequenzen eines zunehmenden Automationsgrads werden insbesondere dann sehr wahrscheinlich, wenn eine kritische Grenze überschritten wird und eine Automation nicht nur informationsanalytische Prozesse übernimmt, sondern auch die aktive Entscheidungsfindung. In Bezug zum *flow chart* Modell (Parasuraman et al., 2000) ermöglichen die Ergebnisse eine Spezifizierung. Der gefundene trade-off bei zunehmendem Automationsgrad zwischen Automationsnutzen auf der einen und -kosten auf der anderen Seite verdeutlicht die Relevanz eines klar formulierten Automatisierungsziels. Nur mit klarem Ziel kann eine Gewichtung der positiven als auch negativen Konsequenzen einer Automatisierung vorgenommen werden. Darüber hinaus ermöglichen die Ergebnisse die Ableitung eines konkreten Leitfadens: Falls die

Aufrechterhaltung manueller Fähigkeiten und des Situationsbewusstseins von besonderer Bedeutung sind, sollten Automationsbestrebungen nicht die Informationsanalyse überschreiten. Falls eine reduzierte Beanspruchung des Operateurs, sowie eine Steigerung der Gesamtleistung wichtiger sind, sollte eine Automation auch Prozesse der Entscheidungsfindung und Handlungsausführung übernehmen.

Um das *flow chart* Modell weiter spezifizieren zu können, fokussierte die darauf folgende Studie (Studie II) das sekundäre Evaluationskriterium der Automationsreliabilität anhand eines Laborexperiments. Ziel der Studie war die Generierung weiterer Erkenntnisse zum Einfluss verschiedener Reliabilitätsniveaus auf die Gesamtsystemleistung, als auch die kognitiven Anforderungen des menschlichen Operateurs. Im Rahmen einer Mehrfachaufgaben-Simulation wurden Probanden durch ein Alarmsystem unterstützt, dessen Reliabilität von 68,75% bis 93,75% variiert wurde. Im Vergleich zu einer rein manuellen Bearbeitung der Simulation profitierten Probanden von der Automationsunterstützung. Die beste Gesamtleistung von Mensch und Automation zeigte sich in Interaktion mit einem Alarmsystem höchster Reliabilität. Wenn die Reliabilität allerdings unterhalb von 70% realisiert wurde, war der zuvor genannte Leistungsvorteil im Vergleich zu den anderen alarmunterstützten Gruppen mit einem stark erhöhten Aufmerksamkeitsaufwand und einer verschlechterten relativen Leistung in einer Parallelaufgabe verbunden. In Anbetracht dieses Gesamtbildes, kann Automation unterhalb einer Reliabilitätsgrenze von 70% nicht als nutzenbringend erachtet werden.

Basierend auf den Ergebnissen der ersten und zweiten Studie, bestand das Hauptziel des letzten Experiments (Studie III) in der Untersuchung möglicher Interaktionseffekte zwischen FA und Automationsreliabilität. Unter Verwendung der gleichen Versuchsumgebung wie in der zweiten Studie, arbeiteten Probanden zusammen mit Automationsunterstützung, welche sowohl in Bezug auf die FA als auch das Reliabilitätsniveau über die jeweilig kritischen Grenzen variiert wurde. Die Ergebnisse zeigten erneut einen starken Einfluss der Automationsreliabilität auf die Aufmerksamkeitsstrategien und Leistung der Probanden. Während relativ zuverlässig arbeitende Automation (Reliabilität über 70%) die Aufmerksamkeitsanforderungen der Probanden verringerte, führte eine Automationsunterstützung welche die Reliabilitätsgrenze verletzte zu keinerlei Vorteilen, weder bezüglich Aufmerksamkeitsanforderungen, noch bezüglich der Leistung. Aufgrund dieser Ergebnisse zum Faktor Reliabilität, kann das sekundäre Evaluationskriterium des *flow chart* Models durch einen konkreten Leitsatz ergänzt werden: Falls eine

Automationsreliabilität über 70% nicht garantiert werden kann, sollte die entsprechende Funktion nicht automatisiert werden!

Allerdings konnte die dritte Studie keinerlei Einfluss der FA auf die Aufmerksamkeitsanforderung und die Leistung der Operateure nachweisen. Eine mögliche Erklärung bezieht sich auf die Realisierung der FA-Faktorausprägung, welche die kritische FA-Grenze in Richtung eines hohen Automationsgrads überschreitet. Während vorherige Studien Automationen miteinander verglichen, die entweder die Informationsanalyse übernahmen (niedriger Automationsgrad) oder die aktive Entscheidungsfindung (hoher Automationsgrad), wurde in dieser Studie als hohe Faktorausprägung eine Automation implementiert, die nicht nur die Entscheidung übernimmt, sondern auch noch die Handlungsausführung. Diese abweichende Operationalisierung verändert die Rolle des Operateurs hin zu einem Supervisor. Als solche, könnten sich Operateure für die Gesamtaufgabe stärker verantwortlich fühlen als Operateure, welche noch immer für die Handlungsausführung entsprechend der Direktiven einer Automation zuständig sind. In diesem Sinne weisen die Ergebnisse darauf hin, dass die Automatisierung einer Gesamtaufgabe eine angemessenere FA darstellen könnte als eine Automatisierung, welche nur die Entscheidungskomponente einschließt. Dieser Argumentation folgend, wird die Berücksichtigung möglicher Interaktionseffekte zwischen FA und Reliabilität bei der Entwicklung automatisierter Systeme im *flow chart* Model ergänzt.

Zusammenfassend bietet die vorliegende Dissertation nicht nur aus theoretischer Sicht neue Erkenntnisse bezüglich des Einflusses von Automationscharakteristika auf die Gesamtsystemleistung und die kognitiven Anforderungen des Operateurs, sondern ermöglicht auch detaillierte Leitlinien, welche in der praktischen Anwendung zu einer effektiveren und effizienteren Automationsentwicklung beitragen können.

Abstract

The objective of this thesis was to gain further insight into impacts of automation on overall system performance and cognitive demands of the human operator. Past research has revealed two important automation characteristics that are crucial to determine if automation support is beneficial or rather deteriorates performance compared with no automation support. One characteristic concerns the function allocation (FA) between human and automation. The second characteristic is the reliability of automation that determines to what extent the operator can rely on the proper functioning of the automated system. With regard to these two aspects, Parasuraman, Sheridan and Wickens (2000) have proposed a flow chart model that attempts to help developers with automation design in terms of an appropriate function allocation between human and automation. After a first consideration of what should be automated, the model suggests primary criteria to evaluate possible consequences of a proposed FA in terms of human performance consequences and imposed cognitive demands. When the initial FA has held out against this primary evaluation, the secondary evaluative criterion of reliability has to be considered for an appropriate FA-decision.

In line with the model's consecutive structure three studies were conducted. The first study is a meta-analysis, which addressed impacts of different FA on Parasuraman et al.'s primary evaluative criteria: operator performance and cognitive demands (Parasuraman et al., 2000). When automation functions properly, results reveal a clear automation benefit for performance and operators' workload with increasing degree of automation (DOA). However, under conditions of automation breakdown increasing the DOA increases the risk of negative consequences in terms of lacking manual performance and situation awareness. Therefore, findings propose that an appropriate function allocation can only serve two of the four aspects. More specifically, negative consequences of automation seem to be most likely when DOA moves across a critical boundary between automation supporting information analysis and automation supporting decision-making. In the context of the flow chart model (Parasuraman et al, 2000) these results provide an opportunity to specify the model. The finding of a direct trade-off between costs and benefits of automation illustrates the importance to consider the objective of automation implementation. Moreover, a concrete guideline for identifying an appropriate FA can be derived from results: If maintenance of skills and situation awareness are crucial, automation should not exceed information analysis. If reduced operator workload and performance benefits are more important, automation should include decision-making functions and action implementation, respectively.

To further specify the flow chart model, the subsequent study focussed on the secondary evaluative criterion, automation's reliability, in an experimental setting. The study's objective was to provide further insight into effects of different levels of reliability on overall system performance, as well as operators' cognitive demands. Within a multi-task simulation, an alarm system differing in reliability from 68.75% to 93.75% supported participants in one of the tasks. In contrast to performing all tasks manually, participants benefited from the alarm support with best performance for the highest reliability condition. However, when reliability was realised below 70% this performance benefit was associated with an increased attentional effort, and a declined relative performance in a concurrent task compared to the other alarm-supported groups. Hence, regarding the overall picture of results, automation below a reliability boundary of 70% cannot be considered as beneficial.

Based on findings of the first and second study, the main objective of the last experiment was to investigate the interaction of FA with the level of reliability. Within the same multi-task simulation as in the second study, automation support differed with respect to FA and reliability, both factors varying across the proposed critical boundaries. Results again revealed a strong impact of automation reliability on participants' attentional strategies and performance. Whereas a fairly reliable automation (above 70%) relieved participants' attentional demands, an automation that violated the reliability boundary was not beneficial, neither in terms of attention nor in relation to performance. Based on these findings regarding reliability, the secondary evaluative criterion of the flow chart model could be specified by a concrete guideline for automation designers: If reliability above 70% cannot be guaranteed, do not automate the function!

However, the third study did not reveal any impact of FA on the reported effects. A possible explanation for this non-finding relates to the realisation of the high-stage automation. Whereas prior studies have compared information automation (low-stage automation) with decision-making automation (high-stage automation), this study implemented high-stage automation by automating the entire task. This changes the operators' role to that of a supervisor. As such, they might feel more responsible for the entire task compared to operators who are still in charge of implementing automation's directives. Therefore, a full automation of a task might represent a more appropriate function allocation, if a high reliability cannot be assured, instead of only automating up to decision-making functions. Accordingly, the flow chart model of automation design was extended with regard to possible interaction effects of FA and reliability.

In sum, this thesis provided not only new theoretical insight into impacts of automation characteristics on overall system performance and cognitive demands of the human operator, but also detailed guidelines that may support practitioners in effective and efficient automation design.

General Comments

All publications used in this thesis are published (study I and study II) or under review (study III) in peer-reviewed journals. Due to differences in journal guidelines the citation styles and reference lists differ between manuscripts. Furthermore, the complete work was written in British English, except study I, which was published in American English due to the journal's guidelines. The thesis is formed by three manuscripts' respective studies.

Abbreviations

AI	action implementation
ANOVA	analysis of variance
AOI	area of interest
AS	action selection
ATC	air traffic control
B	block
DA	decision automation
DOA	degree of automation
e.g.	for example (Latin: exempli gratia)
et al.	and others (Latin: et alii)
FA	function allocation (English) / Funktionsallokation (German)
Hz	Hertz
IA	information automation
IAc	information acquisition
IAn	information analysis
i.e.	that means (Latin: id est)
LSA	loss of situation awareness
M	mean
MABA-MABA	men are better at-machines are better at
MATB	Multi Attribute Task Battery
ms	milliseconds
MTBF	mean time between fixations
NASA-TLX	National Aeronautics and Space Administration-Task Load Index
OOTLUF	out of the loop unfamiliarity
p.	page
RED system	remote eye tracking system
RMSE/ RMS error	root mean squared error

SA	situation awareness
SAGAT	situation awareness global assessment technique
SART	situation awareness rating technique
SD	standard deviation
TCAS	traffic collision avoidance system
TU	Technische Universität
UAV	unmanned aerial vehicle/ unmanned air vehicle
vs.	versus

Table of Contents

Appendix (CD)

1. Introduction

1.1 General Introduction

Automation is all around us. Modern life is not imaginable without highly capable computer systems. No matter if we are driving to a friend in our car with automatic cruise control, if we are going on vacation by plane, or if we are at work interacting with an industrial robot arm: computer-based automation has changed the way we live and work.

The invention of the microprocessor chip in the 1960s enabled the development of highly elaborated automation and still we are not at the top end of this exponentially progressing trend. Today, automation does not remain limited to automated provision of information (e.g. alarm systems) or execution of actions (e.g. tele-robots) but also involves computer-based support of analysing information (e.g. diagnostic assistance in mammography) or choosing appropriate actions (e.g. satnav in cars).

Benefits of increasing automation are manifold. Economic reasons for automation are increased productivity along with cost reduction, but also safety related gains (Satchell, 1998). For example, aviation industry benefited from the implementation of cockpit automation in terms of reduced flight times, increased fuel efficiency, and a more efficient navigation (Billings, 1997; Wickens, Hollands, Banbury, & Parasuraman, 2013; Wiener, 1988). The user of automation, the human operator, has also benefited in terms of more flexibility by relieving operators from several tasks and thereby reducing their workload.

Besides the technological progress and according benefits, there are however still aspects of human-automation interaction that are not very well understood and probably lead to unintended consequences. With the enthusiasm of technological possibilities, automation was seen as the solution to overcome human error and was implemented whenever feasible. However, this technology-centred approach (Sarter, Woods, & Billings, 1997) could not keep up to its promises, which was dramatically shown by numerous (near-)accidents related to problems in human operator's interaction with automated systems in industrial/ professional settings (e.g. partial nuclear meltdown at Three Mile Island in 1979, Bophal gas tragedy in 1984, grounding of the cruise ship Royal Majesty in 1995) or daily life (e.g. numerous anecdotal reports of drivers who followed their navigation system and entered wrong streets or drove into rivers that were clearly identifiable as rivers).

Automation designers had and still have to recognise that automation should not be implemented just because it is possible. As Sheridan and Parasuraman (2005) state: "[…] to engineer the automation and expect the human to accommodate to it can be a recipe for disaster." (p. 94). Therefore, a human-centred approach (Billings, 1991) that focuses on the interaction of humans with automation is mandatory to ensure an appropriate use and in consequence a safe operation of system (Wickens, Mavor, Parasuraman, & McGee, 1998). This focus is leading in human factors research and the underlying notion of the current work. It is important to gain insight into which factors affect human-automation interaction. A large body of research suggests that two aspects of automation are particularly crucial in affecting overall performance and human operators' cognitive demands: function allocation between human and automation, as well as automation's reliability (Cummings & Mitchell, 2007; Endsley & Kiris, 1995; Goddard, Roudsari, & Wyatt, 2012; Kaber, Onal, & Endsley, 2000; Layton, Smith, & McCoy, 1994; Lorenz, Di Nocera, Röttger, & Parasuraman, 2002; Manzey, Reichenbach, & Onnasch, 2012; Parasuraman, Molloy, & Singh, 1993; Wickens, Dixon, Goh, & Hammer, 2005; Wickens, 2000).

With regard to the two aspects, Parasuraman, Sheridan and Wickens (2000) have proposed a flow chart model that serves as an orientation for automation decisions. Impacts of function allocation on human/ overall system performance as well as on operators' cognitive demands are defined as primary evaluative criteria for automation. Reliability is considered as a secondary evaluative criterion. The consecutive structure of the model accounts for the fact, that (un-)reliability may have differential effects as a function of more or less automation. First of all, function allocation between human and automation has to be determined in order to be able to evaluate effects of reliability of the chosen form of automation.

The current work comprises three studies dealing with the impact of function allocation and automation's reliability on human-automation interaction. The levels and stages taxonomy as well as the flow chart model by Parasuraman et al. (2000) serve as theoretical framework. The models will be described in detail in chapter 1.3.

The first study addresses effects of function allocation on operator performance and cognitive demands in form of a meta-analysis. The second study was conducted as a laboratory experiment and focusses on human adaptation to automation's reliability in terms of attention allocation, an indicator for cognitive demands, and performance. The third study combines findings of the latter two and varied both factors, function allocation and reliability, at the same time in an experimental setting to gain insight into differential effects on performance and attention allocation.

Introducing the three studies, the following sections provide basic notions concerning automation and human-automation interaction. Subsequently, concepts of function allocation and automation reliability as well as possible interaction effects of both factors are discussed as aspects directly related to automation. The two concepts constitute the main influence on human-automation interaction that have to be considered regarding an appropriate interaction with automated systems.

1.2 Automation – a Definition

Wiener and Curry (1980) draw two extreme pictures that underline wishes and fear, which are associated with automation. When talking about automation one may consider it as the technological savior of society relieving humans from unpleasant work and eliminating human error. At the contrary, automation may be seen as a "*collection of tyrannical, self-serving machines, degrading humans, reducing the work force, bringing wholesale unemployment, and perhaps even worse, offering an invitation to a technological dictator to seize power and build a society run by Dr. Strangeloves, aided by opportunistic, cold-hearted computer geniuses.*" (Wiener & Curry, 1980, p. 995). Both ideas are clearly exaggerated and 34 years later automation has neither enslaved us, nor has it freed us from all discomfort. Nevertheless, automation has simplified work in many domains by taking over so-called 3-D tasks - dull, dirty, and dangerous tasks that are very monotonous, are physically hard work, or cause high amounts of mental workload (Nof, 2009).

The advance of automation first started in the manufacturing domain. In the 1950s D.S. Harder (vice-president of the Ford Motor Company) described automation as a philosophy of manufacturing in which mechanical, hydraulic, or electronic devices should replace human work (Salvendy, 1997). This understanding of automation was however mainly related to physical human work. With the development of more powerful computers the application of automation not only resumes simple action implementation but also comprises cognitive processes like information analysis and complex decision-making. Accordingly, automation can be described as delegation of various functions from a human to a machine agent (Billings, 1991; Parasuraman & Riley, 1997; Raouf, 1988). When the reallocation of functions is irreversible, i.e. a human operator cannot fulfil the task anymore if desired or wanted, this technical system is rather understood as a technical machine than an automation. This dynamic definition of automation is summed up by Parasuraman and Riley (1997): "Today's automation could well be tomorrow's machine." (p. 231). Starter motors for cars are an example for a technical system in which the reallocation of functions is complete and permanent; a redistribution of the task is not possible anymore.

Taking these aspects into account Moray, Inagaki, and Itoh (2000) present a more precise definition of automation, which is adopted in this work. The allocation of functions from the human to automation is differentiated according to the (simplified) human information processing model (Wickens et al., 2013). In this respect, automation is "any sensing, detection, information-processing, decision-making, or control action that could be performed by humans but is actually performed by machine." (Moray et al., 2000, p. 44). Based on this definition, automation can refer to

- the mechanised sensing of environmental variables by artificial sensors,
- data processing and decision-making by computers,
- any mechanical action by devices that apply forces on the environment like motors,
- or communicate processed information to a human operator (Sheridan & Parasuraman, 2005).

An alternative description categorises automation due to its purpose (Wickens et al., 2013). According to Wickens et al. (2013) the first category contains automation of tasks that humans should not perform because of high risk. In domains like nuclear power, for example, complex mathematical operations have to be performed to control certain processes that are too complex for a human operator to do online. Therefore, those tasks are delegated to automated systems. Other examples concern work in hazardous or dangerous environments, like the use of reconnaissance and defusing robots in bomb squads.

The second category deals with automation that overcomes human performance limitations. In this case, humans could do the tasks yet poorly or at the expense of high cognitive/ physical demands. A most recent example for this category is the development of complex unmanned aerial vehicle (UAV) networks in the military domain. Human operators could not sufficiently monitor and delegate multiple UAVs. Therefore, expert systems take over large parts of the monitoring and alert operators when to shift attention to certain tasks or UAVs.

The third category describes automation as an augmentation or assistance of human performance. At first sight, similar to the aforementioned category, the emphasis is on assisting in contrast to resuming parts of the main task. Examples from everyday life are satnav or brake booster support in automobiles.

The fourth and last category focusses on the purpose of automation to increase productivity and decrease costs at the same time.

Automation that serves purposes of category two and three, i.e. automation that assists human operators, is of particular interest for the current work. While these kinds of automation support the human by resuming certain functions of task fulfilment, they also create new challenges that relate to the human-automation interaction. First, the interaction with an automated system represents a new task to the human operator that has to be coordinated in order to be able to benefit from support. Second, with the implementation of automation the human operator is (partly) taken out-of-the-loop from active task fulfilment and rather serves as a supervisor of automation. These new management tasks produce new attentional demands, which, in the worst case, can even exceed the original cognitive demands that were intended to be reduced by automation (Wiener, 1989).

Therefore, human-automation interaction is of special interest in automation research, as an inappropriate interaction may counteract the proposed benefits. In consequence, an elaborated examination should not focus exclusively on automation but on the joint human-automation system as a cooperative team. This approach will be followed in the rest of this work.

1.3 Human-Automation Interaction

The current work's objective is to gain further insight into how automation characteristics impact overall system performance (human and automation) and cognitive demands of the human operator. For that purpose, it is necessary to provide the basic concept of human-automation interaction as a framework in order to understand more specific aspects of this collaboration.

Introducing automation into a complex system does not simply supplant human activity, but rather changes the role of the human operator. In most circumstances, automation changes the role of the human operator from active involvement to passive control. Specific (if not all) tasks are resumed by the automated agent. The human operator serves as automation's supervisor who only intervenes, if the automated system does not function properly (Sheridan & Verplank, 1978).

The concept of supervisory control derived from the characteristics of a supervisor interacting with subordinate human staff members. A supervisor provides general directives that are translated into detailed actions by staff members. The supervisor in turn receives summarised information about process results that are important for further steps from the subordinates. This form of interaction is transferable to human-automation interaction. In particular, Sheridan and Verplank (1978) describe five functions of the human supervisor that have to be performed in consecutive order for optimal interaction with the subordinate automation.

Planning: The operator has to plan the task off-line. This first step normally does not constitute an interaction. However, an initial appropriate mental model of the task and of the accordant automation is necessary to plan which tasks should be performed by automation.

Teaching: Secondly, the task strategy has to be taught to the automation by input and control actions.

Monitoring: Thirdly, the operator has to monitor the automated execution of task fulfilment. Monitoring means allocating attention to different information displays or alternative information sources to verify the desired outcome.

Intervening: In case of deviating outcomes, the operator must intervene to regain control over the process and thus to prevent unintended consequences.

Learning: Sheridan and Verplank (1978) further describe learning from experience in interaction with automation as the fifth human function that constitutes an out-of-the-loop human task that feeds back into the first step: a new planning of task fulfilment.

The third and the fourth function (monitoring, intervening) are of special interest because monitoring automation captures the greatest part of operator work and both, monitoring and intervening, are seen as the most critical parts in direct interaction with automation (Bainbridge, 1983). The main concern with these two operator tasks is that they imply some inconsistences that were first described in Bainbridge's ironies of automation (1983). Concerning monitoring, the human's task is to supervise functions that were quite often automated to prevent human error. However, instead of preventing, human error is just shifted by automation to another position, from the operator to the designer of automation (Parasuraman & Riley, 1997). In addition, the operators' task to intervene and to regain manual control in automated systems is only required in situations that are extremely critical. After a prolonged time of being passive and out-of-the-loop (monitoring task), the human shall resume manual control in situations that automation cannot manage and that are in most circumstances time critical and very complex (Bainbridge, 1983). Bainbridge pinpoints this notion: "By taking away the easy parts […], automation can make the difficult parts of the human operator's task more difficult.” (p. 777).

These ironies of automation are strongly related to the function allocation between human and automation; i.e. to what extent the human is kept in-the-loop of the overall task. Accordingly, costs and benefits of automation have to be considered as a function of more or less automation. Therefore, function allocation presents a central and primary issue in describing human-automation interaction. Consequently, consequences of function allocation on performance and operators cognitive demands are one of the research objectives of the current work, which will be discussed in the subsequent subsection.

1.3.1 Function Allocation between Human and Automation

The question, which functions should be automated, and which should remain with the human operator, emerged since automation has conquered the work place. In the beginning this problem seemed to be easy to answer: Automate everything that is technically feasible; negative consequences of this approach (e.g. skill degradation, loss of situation awareness) have shown that there is no simple answer (Dekker & Woods, 2002).

As one of the first, Fitts (1951, as cited in Sheridan, 2000) suggested a differentiated answer to the question of function allocation. He proposed a list comparing “what men are better at” and “what machines are better at” (MABA-MABA) to underline the strengths and weaknesses of humans and machines as a guideline for assigning functions to one or another. However, there are several concerns with this approach. First: it is doubted that men and machine are comparable. Comparability would imply that humans are equivalent to and completely exchangeable with machines (Jordan, 1963). Second: as technological feasibility has evolved, the Fitts’ list is not very useful today. If we decided on function allocation comparing what men are better at and what machines are better at today, only few tasks would remain with the human operator. However, following this left-over principle (Hollnagel & Bye, 2000) does not consider the human role in the resulting automated system. Third: the competitive idea behind this comparison is not constructive in promoting a collaborative human-automation interaction to designers or engineers (Billings, 1991). And fourth: because of the manifold contexts of automation a list as a guideline is too simplistic for automation development as the context is always important and may change certain decisions which functions to automate. For example, design of automation should differ depending on whether it is implemented in a nuclear power plant or a manufacturing site producing toothbrushes.

Therefore, function allocation models developed later than the Fitts' list went one step back in the attempt to find a standardised categorisation of automated systems with respect to function distribution instead of giving instant guidelines (e.g. Endsley & Kaber, 1999; Endsley & Kiris, 1995; Milgram, Rastogi, & Grodski, 1995; Parasuraman et al., 2000; Riley, 1989; Sheridan, 2000; Wickens et al., 1998).

Common to all these function allocation models is the assumption that automation does not exist in an all-or-none fashion, but rather constitutes a continuum from no support to full automation of all functions. Unfortunately, the fragmentation of this continuum seems to be the crux, as all models suggest different gradations. Whereas some taxonomies suggest five subsequent stages of more or less automation (Endsley & Kiris, 1995; Endsley, 1987), others propose seven (Billings, 1991), ten (Endsley & Kaber, 1999; Sheridan & Verplank, 1978), twelve (Riley, 1989), or even 17 gradations (Bright, 1958). Furthermore, not all taxonomies take into account all possible kinds of automation. Whereas Endsley's and Kaber's ten-stage solution describes tasks that incorporate cognitive as well as action implementing functions (Endsley and Kaber, 1999), Sheridan and Verplank apply their ten stages to higher-level cognitive functions only, therefore reaching a finer grained differentiation (Sheridan and Verplank, 1978). The problem is the differing viewpoints of researchers. Models developed for cognitive tasks, for example, do not apply for information acquisition functions and vice versa. Therefore, the objective to find a standardised categorisation of automation is not yet fulfilled by these models.

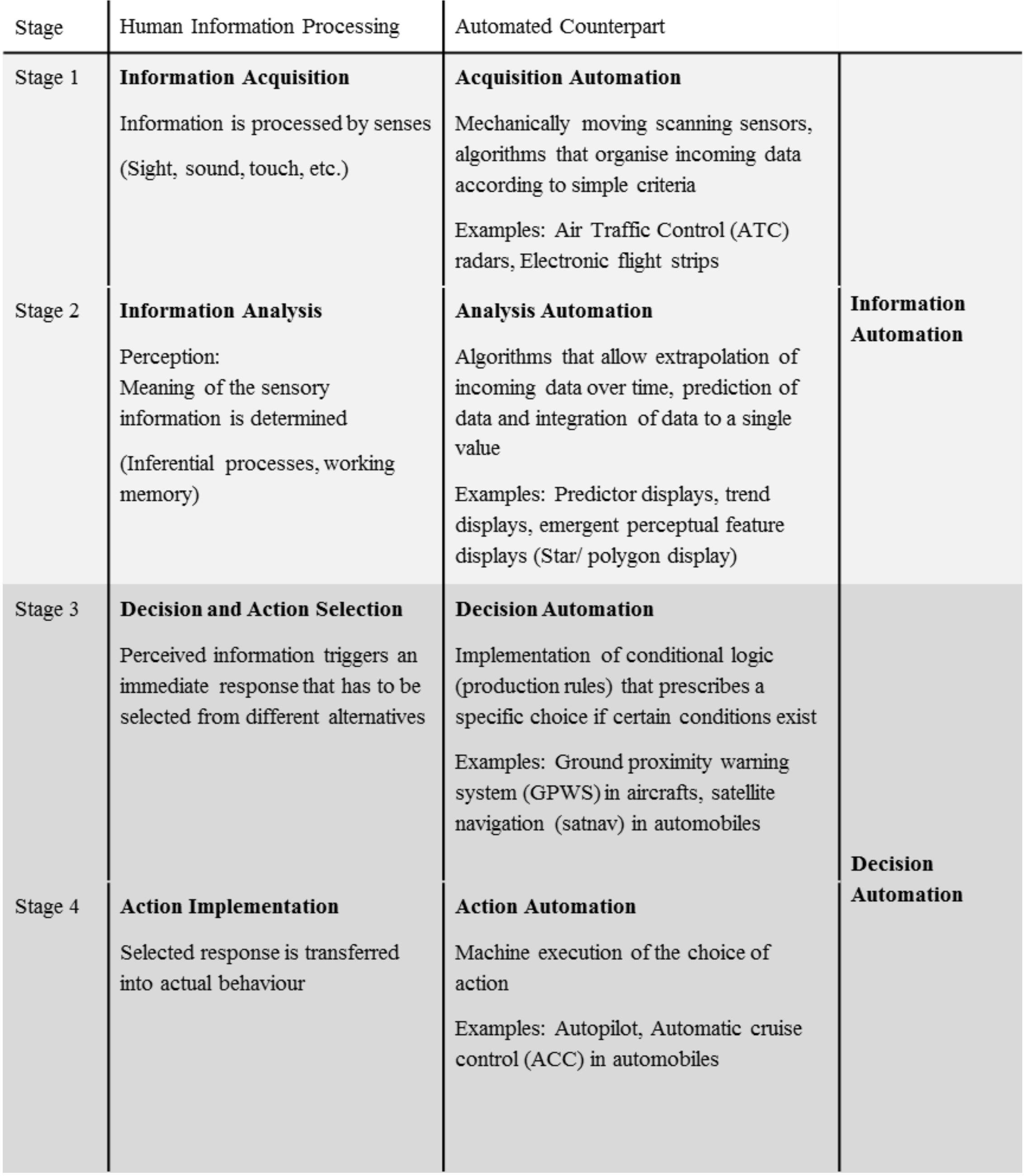
A two-dimensional taxonomy introduced by Parasuraman et al. (2000) adopts the idea of different levels of automation and further provides a stage component that allows for integrating different functions automation can apply to. The first dimension of this framework adopts Sheridan's and Verplank's ten-level scale of higher cognitive functions (1978). The dimension ranges from level 1: the computer offers no assistance, the human is in full responsibility of the task; to level 10, in which the computer decides everything and acts autonomously without a possibility of the human to intervene (Sheridan & Verplank, 1978). Between these two extremes eight consecutive levels are defined referring to more or less human involvement. The ten-level scale is applicable to output functions, i.e. automation of decision selection and action implementation.

However, there are more functions that can be automated, like input functions, in which information is sampled. This functional component of automation is covered with the second dimension: stages of automation (Parasuraman et al., 2000). The stage dimension is adopted from Wickens' model of human information processing (Wickens et al., 2013). This simplification of human information processing provides a useful structure for describing certain principles that are important to human-automation interaction. In order to comprehend the transfer of this model to the stage component of automation, the human information processing model is described first, followed by a description of the transfer to Parasuraman et al.'s (2000) stage component.

The human information processing model contains four stages. At stage 1; information acquisition; information is gathered from the environment. This stage incorporates processes on a sensing level that can be described as a pre-processing of information prior to conscious perception. At stage 2; information analysis; the retrieved data is actually perceived and further manipulated in the working memory. This also implies certain cognitive operations like rehearsal, integration, inference. At stage 3; decision and action selection; decisions are made based on previous processing of the retrieved information. At stage 4; action implementation; an action is initiated that transfers the decision to actual behaviour (Wickens et al., 2013).

Parasuraman et al. (2000) adopt these four stages to describe automation of different human information processing functions. Table 1 provides a comparison of the stages of human information processing and the equivalent automation stage.

Table 1. Comparison of stages of human information processing (Wickens et al., 2013) and the equivalent automation stage according to Parasuraman et al. (2000)



Stage	Human Information Processing	Automated Counterpart	
Stage 1	**Information Acquisition** Information is processed by senses (Sight, sound, touch, etc.)	**Acquisition Automation** Mechanically moving scanning sensors, algorithms that organise incoming data according to simple criteria Examples: Air Traffic Control (ATC) radars, Electronic flight strips	**Information Automation**
Stage 2	**Information Analysis** Perception: Meaning of the sensory information is determined (Inferential processes, working memory)	**Analysis Automation** Algorithms that allow extrapolation of incoming data over time, prediction of data and integration of data to a single value Examples: Predictor displays, trend displays, emergent perceptual feature displays (Star/ polygon display)	
Stage 3	**Decision and Action Selection** Perceived information triggers an immediate response that has to be selected from different alternatives	**Decision Automation** Implementation of conditional logic (production rules) that prescribes a specific choice if certain conditions exist Examples: Ground proximity warning system (GPWS) in aircrafts, satellite navigation (satnav) in automobiles	**Decision Automation**
Stage 4	**Action Implementation** Selected response is transferred into actual behaviour	**Action Automation** Machine execution of the choice of action Examples: Autopilot, Automatic cruise control (ACC) in automobiles	

As stages 1 and 2 represent automation that is related to input functions, both stages can be summarised as information automation. Stages 3 and 4 refer to output functions and can be summarised as decision automation (Parasuraman et al., 2000).

Due to the two-dimensionality of the framework, a system can be characterised according to the functions that are automated (stage dimension) and the level of automation on each stage (level dimension). With this approach every kind of automation can be described and compared in a standardised manner, including all kinds of functions. This outmatches other models that are only applicable to a certain kind of automation (e.g. Endsley & Kiris, 1995; Milgram, et al., 1995; Riley, 1989; Sheridan, 2000).

A limitation of Parasuraman et al.'s model (2000) relates to the level-dimension, which is adopted from Sheridan's and Verplank's ten-level taxonomy of automation (1978). This taxonomy can be easily applied to stage 3; decision and action selection; but is not transferable to the other stages of automation. Nevertheless, the idea that different stages can be automated to different levels is highlighted by this framework and useful to describe automation in detail.

As an example, consider two health care automated systems, which are characterised and compared regarding the amount of automation. System A may (1) alert abnormal patient symptoms, and (2) integrate these symptoms to form an intelligent diagnosis of the patient condition. System B may further (3) recommend a treatment based upon the diagnosis as well as (4) carry out the action when approved by the human. Both systems are illustrated in figure 1.

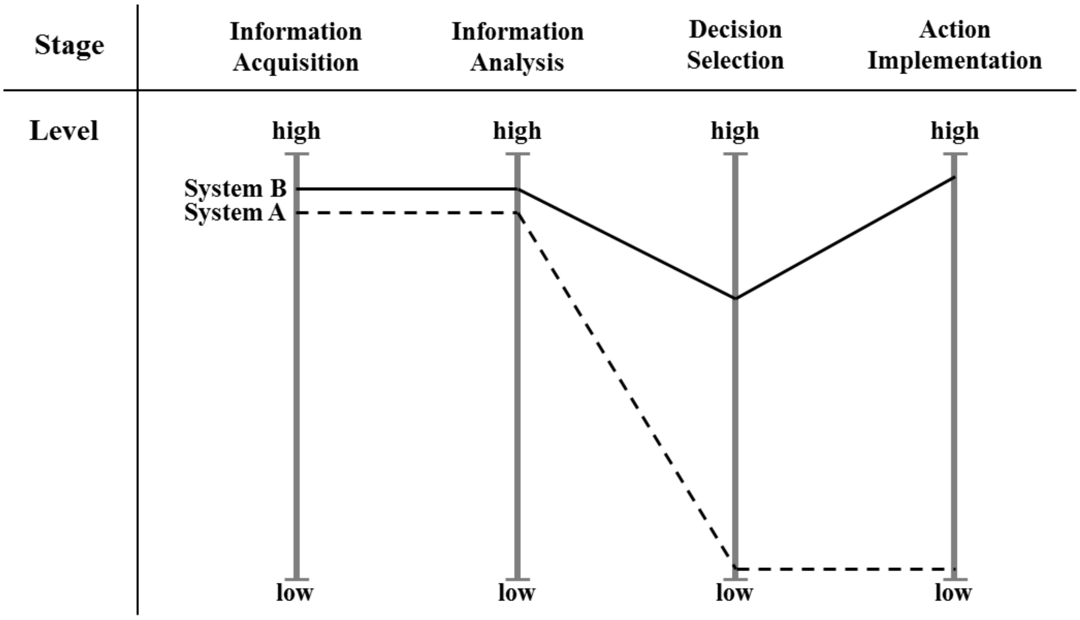


Figure 1. Characterisation of automated systems A and B, applying the stages and levels taxonomy proposed by Parasuraman et al. (2000). A detailed description of the systems is provided in the text. Note that for Information Acquisition and Information Analysis both systems have the same characteristics. For clarity the lines are organised one below the other.

By applying the stages and levels model (Parasuraman et al., 2000), it is possible to directly compare systems and highlight the commonalities and differences of systems. Furthermore, possible costs and benefits of automation can be directly related to specific levels or stages and not only to more or less automation. Therefore, causes for specific consequences can be identified. For example, the automation of decision-making, stage 3, may result in a loss of situation awareness (Endsley, 1995, 1996), whereas stage 2 automation may not negatively affect situation awareness.

Based on the levels and stages model, Parasuraman et al. (2000) further developed a flow chart model that attempts to help developers with automation design in terms of an appropriate function allocation between human and automation. The flow chart model builds the structural framework of this thesis and is depicted in figure 2.

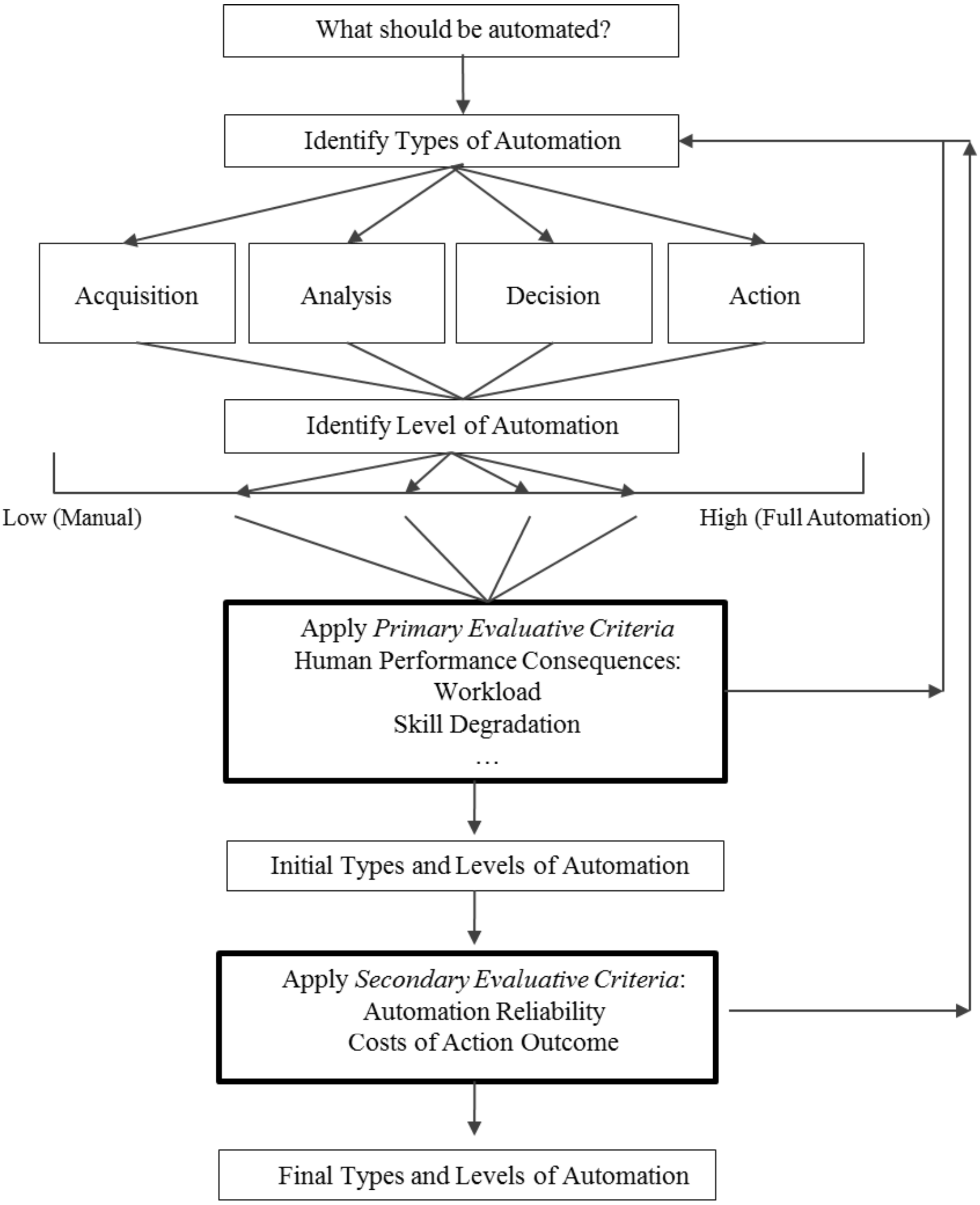

What should be automated?
Identify Types of Automation
Acquisition
Analysis
Decision
Action
Identify Level of Automation
Low (Manual)
High (Full Automation)
Apply *Primary Evaluative Criteria*
Human Performance Consequences:
Workload
Skill Degradation
…
Initial Types and Levels of Automation
Apply *Secondary Evaluative Criteria*:
Automation Reliability
Costs of Action Outcome
Final Types and Levels of Automation

Figure 2. Flow chart model of automation design (Parasuraman et al., 2000)

After a first consideration of what should be automated according to stages and levels, the primary criteria evaluate possible consequences of a proposed function allocation in terms of human performance consequences and imposed cognitive demands.

Primary evaluative criteria (upper bold box, figure 2): For this evaluation, two conditions have to be differentiated: automation under normal-operating conditions and under conditions of automation breakdown.

Under normal-operating conditions human operators mostly benefit from automation. Considering the progress of technology, human operators are confronted with increasing task demands. Automation can considerably reduce those task requirements by taking over certain parts, thus reducing operators' workload (e.g. Breton & Bossé, 2003; Sheridan & Parasuraman, 2005; Sheridan, 2002). This is particularly true for higher-stage automation that does not only resume input functions but also cognitive tasks related to decision-making and action selection. Furthermore, performance of the joint human-automation system normally exceeds unaided performance. This is also true for concurrent task performance as automating certain functions reduces operators' attentional demands to the automated task. Consequently, freed cognitive resources can be reallocated to concurrent tasks that additionally benefit from automation (Breton & Bossé, 2003; Pritchett, 2001).

Under conditions of automation breakdown, however, negative consequences of automation are likely, which may result in catastrophic failures. Negative effects are directly related to the operator's out-of-loop performance problem (Wickens, 2000), which is more likely in interaction with a higher-stage automation compared to less capable automation. When human operators are not anymore actively involved in task completion but serve as passive monitors, they are frequently slow in detecting problems that need a manual intervention. Once the failure is detected, additional time is needed to determine the system state and to understand why the automation failed. These problems are associated with an automation-induced loss of situation awareness. Situation awareness (SA) is defined as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future (Endsley, 1988, p. 97). Endsley further decomposes the concept of SA into three hierarchical levels (Endsley, 1995). On level 1, SA is equitable with the first part of the definition, the perception of elements in the environment including their status, attributes, and dynamics. For example, when driving a car, one needs to know where other vehicles and obstacles are that are relevant to one's own vehicle.

Level 2 describes the comprehension of the current situation and is based on information obtained at level 1 SA. Comprehension goes beyond simply being aware of the information obtained but rather includes an understanding of the relevance of information to gain an overall picture of the situation. For example, a driver needs to put together information of decelerating vehicles around him/ her and traffic lights that are in near distance to understand the behaviour of other road users.

Level 3 SA focuses on the ability to project near future actions of the elements in the environment. This last step is only accomplishable on the basis of level 1 and level 2 SA. Referring to the example, based on the driver's understanding that other road users are decelerating because the traffic light has switched from green to orange the driver is able to infer that the other vehicles attempt to stop at the traffic light.

SA therefore plays an important role in interacting with the environment, which also includes the interaction with automation. SA is critical to effective decision-making resulting from an appropriate monitoring and understanding of the environment (Endsley, 1996).

Another major concern introducing automation is the potential loss of manual skills. When task execution is completely delegated to automation under normal-operating conditions, manual skills are likely to deteriorate with lack of use (Endsley & Kiris, 1995; Wiener, 1989). However, these skills are not only crucial in operating a system manually in cases of automation breakdown, but also for detecting an automation malfunction and consequently the need to intervene manually (Endsley & Kiris, 1995).

Furthermore, automation under conditions of breakdown may also negatively affect operators' workload. While workload should be reduced by automation under normal operating conditions it can be even intensified under abnormal system state by automation. Wiener subsumes this imbalance in which automation reduces workload in routine, low-workload situations, but exacerbates even more workload during busy, high-criticality, event-driven operations as clumsy automation (Wiener, 1988). Clumsy automation is a form of poor coordination between human and machine. The reason for clumsiness is a lacking adaptation of automation to the activity phases of the human operator that normally differ depending on context factors like process state. For pilots workload differs in dependence of flight phase, for example.

Considering these costs but also the benefits of automation in the design process should lead to a more deliberated function allocation decision. However, every implementation of automation is still a single-case decision. Even with a large body of research regarding effects of function allocation on operator performance and cognitive demands, findings are still mixed thereby making a deduction of more general principles not possible (e.g. Endsley & Kiris, 1995; Kaber et al., 2000; Layton et al., 1994; Lorenz et al., 2002; Manzey et al., 2012; Sarter & Schroeder, 2001). One reason for this lack of generalisation is due to the fact, that experimental research does not always follow the taxonomy proposed by Parasuraman et al. (2000) and has evaluated numerous forms of automation in diverse contexts. Furthermore, results of single studies are inconclusive and suffer from a limited statistical power (e.g. Kaber et al., 2000; Lorenz et al., 2002; Sarter & Schroeder, 2001). Thus, a more valid overall picture to allow conclusions about which level or stage of automation comes with certain benefits or costs is needed.

Therefore, the objective of the first study in this work is to provide an overall picture by quantitatively combining single studies of function allocation in a meta-analysis. The study examines effects of function allocation on performance, as well as cognitive demands under routine and automation breakdown conditions.

1.3.2 Automation Reliability

Besides the primary evaluative criteria, consequences of function allocation, other aspects have to be further considered that may relativise decisions that are based on the primary evaluation (consequences of function allocation). In the flow chart model, these are summarised as the next hierarchical step, the **secondary evaluative criteria** (lower bold box in figure 2). One secondary criterion is concerned with the economic consequences of automation breakdown, which is not addressed in this work (see Parasuraman et al., 2000). Another criterion is directly related to automation design and is discussed in detail in this chapter: the reliability of automation.

In contrast to automation breakdown, in which automation does not function at all and cannot be used anymore, reliability describes the performance of automation. Because of imperfect sensors, algorithms, as well as the challenge to interpret a noisy and uncertain world, automation may not always be right in terms of a given alarm, a proposed diagnose or an executed action. Furthermore, automation may not function properly because certain events are missed, i.e. automation may fail to give an alarm, fail to provide a diagnosis, or miss to implement a required action (Parasuraman et al., 2000). Depending on the realised stage of automation, reliability can therefore be defined as the proportion of correctly indicated critical events (information automation), correctly given diagnoses, suggested decisions, or correctly executed actions (decision automation) divided by the total number of operations in the automated task.

Reliability is one of the most important perceivable characteristics, as it affects human monitoring of an automated system as well as monitoring of alternative information (Lee & See, 2004; Lee & Moray, 1992; Muir, 1987). When reliability is high, operators are relieved from continuous monitoring, as they can depend on the proper functioning of automation. For example, if we trust our satnav we might not screen all road signs when driving on the freeway, because we depend on the proper functioning of it, and that it will lead us to our desired destination. However, when reliability is not sufficient, more cognitive resources have to be allocated to the automated task in order to compensate for automation's imperfection. For the driving example this would imply that we had to screen all road signs to verify proposed routes of the satnav.

Thus, to what extent operators depend on automation is a function of automation's reliability, i.e. the higher the reliability, the more operators depend on it and vice versa. However, this association of reliability and dependency is not always appropriate. When reliability is high, humans tend to go the way of least cognitive effort and monitor automation even less than would be required. This misuse of automation (Parasuraman & Riley, 1997) has been the objective of a large body of research (e.g. Bagheri & Jamieson, 2004; Bahner, Hüper, & Manzey, 2008; Bailey & Scerbo, 2007; Endsley & Kiris, 1995; Goddard et al., 2012; Kaber & Endsley, 1997; Manzey et al., 2012; Singh, Tiwari, & Singh, 2009).

When interacting with automation in the context of supervisory control, i.e. automation that has to be monitored continuously by operators, the tendency to overly rely on and subsequently not monitor automation is subsumed under the concept of complacency (Parasuraman & Manzey, 2010). Whereas there is still no consensus on the definition of complacency, Parasuraman and Manzey (2010) point out three core features that underlie different definitions (e.g. Billings, Lauber, Funkhouser, Lyman, & Huff, 1976; Wiener, 1981), as well as the operationalisation of complacency in most of the experimental studies (e.g. Bahner, Hüper, & Manzey, 2008; Bailey & Scerbo, 2007; Manzey et al., 2012; Parasuraman et al., 1993). The first feature relates to the fact that operator monitoring of automation, like in supervisory control, has to be involved in the task setting. Secondly, the frequency of operator monitoring has to be lower than some predefined optimal value. The third feature requires that there has to be some direct observable negative effect on system performance due to the suboptimal monitoring of automation (Parasuraman & Manzey, 2010). If these features apply, an operator's monitoring can be classified as not sufficient and therefore the operator's adaptation to automation is disproportionate. The association between automation's reliability and the operator's dependency on automation is biased. In the worst case, this under-monitoring of automation may have dramatic consequences, when the human operator does not detect automation failures (or detects them too late).

In interaction with decision support or diagnostic automation, two distinct forms of operators' dependency on automation can further be differentiated: Compliance refers to an operator's response, when automation indicates that an action is needed. In contrast, reliance describes an operator's tendency to rely on automation, when it indicates that the monitored process runs properly and the operator accordingly does not have to take any action (Meyer, 2001, 2004). Pilots, for example, may always comply with the traffic collision avoidance system (TCAS) and directly take evasive actions without cross-checking when TCAS indicates a possible hazard. However, pilots' reliance may not be that strong, so that they monitor the airspace for possible hazardous traffic concurrently to detect any possible miss by TCAS.

When operators overly rely on or comply with decision support or diagnostic automation, Mosier and Skitka (1996) describe this misuse of automation as automation bias, a term explicitly related to other cognitive biases like the availability heuristic (Tversky & Kahneman, 1973). Referring to other cognitive biases, Mosier and Skitka (1996) define automation bias as "the tendency to use automated cues as a heuristic replacement for operators' vigilant information seeking and processing […]" (p. 205). If automation functions properly, this cognitive shortcut is very efficient in terms of reduced cognitive workload. However, if automation does not function perfectly, an overreliance may lead to omission errors: errors that are related to misses of the automation and result if human operators do not take evasive action because they were not informed by automation. Another possibility for failure results from overcompliance with automation. In this respect, commission errors result, when automation produces a false alarm or diagnosis and operators inappropriately follow information or directives provided by automation. This tendency is observable even if alternative information clearly contradicts the automated advice (Mosier, Skitka, Heers, & Burdick, 1998; Mosier & Skitka, 1996).

Contrary to operators' adaptation problems in interaction with highly reliable automation, there is also evidence that operators are particularly worse in adapting to automation with low reliability (e.g. Dixon, Wickens, & Chang, 2004; Dixon & Wickens, 2006; Wickens et al., 2005; Wickens & Dixon, 2007). In a review article, Lee and See (2004) describe experimental findings from trust research regarding human-automation interaction. Below a certain reliability level trust declines quite rapidly and is not associated with an actual reliability level anymore (Lee & See, 2004). Estimates of this reliability level range from 90% (Moray et al., 2000) to 60% (Fox, 1996). Furthermore, a more recent quantitative literature review conducted by Wickens and Dixon (2007) relates automation’s reliability to the joint human-automation performance. This review suggests a reliability boundary around 70%. Whereas reliability above this level means better joint human-automation performance, automation less reliable than 70% even yields worse performance compared to working with no automation. Thus, effective compensation for unreliability seems to be possible to a certain level only.

Because of the behavioural consequences that reliability imposes, it is important to further understand its impact and how people adapt to differing reliability levels. In sum, studies conducted thus far show an inconsistent pattern of results (e.g. Bagheri & Jamieson, 2004; Bahner, Hüper, & Manzey, 2008; Bailey & Scerbo, 2007; Dixon, Wickens, & Chang, 2004; Dixon & Wickens, 2006; Singh, Tiwari, & Singh, 2009). However, most of these studies only compared relatively extreme reliability levels and missed to describe the characteristics of operators’ adaptation to automation across a more complete range of reliability. The assumption of a reliability boundary is therefore only an estimated cut-off value, but what we do not know is how this change from supportive automation to useless automation develops, i.e. how humans adapt to reliability values around the proposed boundary by Wickens and Dixon (2007). Furthermore, most studies conducted thus far, evaluated effects of automation reliability only via performance measures. Effects of reliability however are already likely at a preceding level: the operators` cognitive demands like attention allocation. As was already proposed by Parasuraman and Manzey (2010) in the context of complacency, consequences of reliability should not only be evaluated by performance measures but also via cognitive variables.

Therefore, the objective of the second study is to provide further insight into effects of different levels of reliability on operators’ adaptation strategies with respect to overall system performance and operators’ cognitive demands.

1.3.3 Function Allocation and Automation Reliability

Referring to Parasuraman et al.’s (2000) flow chart model of automation design (see Figure 2), reliability represents the second evaluative criterion, i.e. the evaluation of reliability always depends on a foregoing decision of function allocation. This structure implies that (un-)reliability may have different consequences for different stages or levels of automation. Even with an elaborated understanding of impacts of function allocation or reliability, the combination of both factors may lead to not foreseen effects. Therefore, Parasuraman et al. (2000) demand that experimental studies evaluating effects of different function allocations for operator performance and cognitive demands should vary function allocation and reliability concurrently. However, only a limited number of studies have become available that collected empirical data on possible interaction effects of both factors and results are inconsistent (Crocoll & Coury, 1990; Galster & Parasuraman, 2004; Galster, 2003; Rovira, McGarry, & Parasuraman, 2007; Rovira, Zinni, & Parasuraman, 2002; Sarter & Schroeder, 2001). Some of these studies suggest an interaction effect between function allocation and reliability. In particular, it was found that same levels of unreliability led to worse effects on human performance in case of decision versus information automation (e.g. Crocoll & Coury, 1990; Rovira et al., 2007; Sarter & Schroeder, 2001). However, other studies report detrimental effects of unreliable automation already for information automation or even worse performance for information compared to decision automation (Galster & Parasuraman, 2004; Galster, 2003). Furthermore, most of these studies are afflicted with some limitations concerning aspects of experimental design. Firstly, most of the studies only compared perfectly reliable automation trials with trials of automation breakdown (Crocoll & Coury, 1990; Rovira et al., 2002; Sarter & Schroeder, 2001). However, such a comparison of extremes, perfect reliability vs. automation breakdown, does not allow conclusions about effects of different levels of reliability in interaction with varying function allocation.

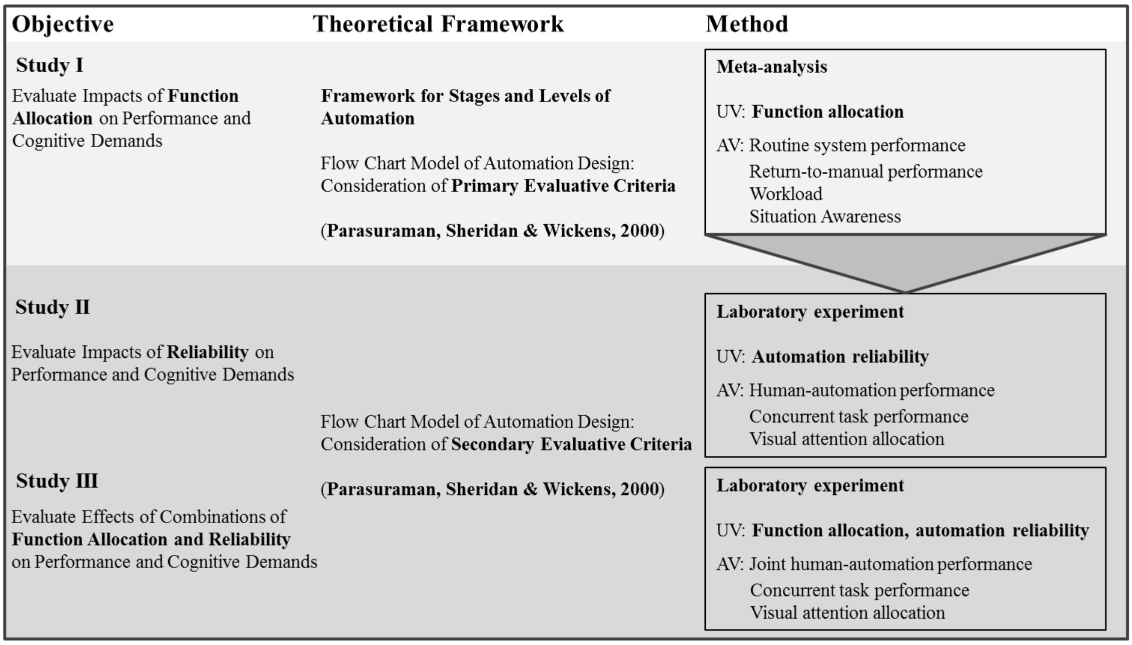
Secondly, most of the previous research missed to consider effects on operators' cognitive demands induced by different sorts of automation. However, benefits of automation as well as unintended costs are often related to operators' cognitive demands. As Moray and Inagaki (2000) have pointed out, an exclusive consideration of performance data may result in biased interpretations. For example, when the detection rate of automation failures by participants is low, it is often claimed that participants are complacent (e.g. Parasuraman et al., 1993). However, as long as attention allocation was not additionally measured, this interpretation is questionable because nothing is known about the frequency with which participants actually monitored the task (Moray & Inagaki, 2000). The reason for not-detecting or not reacting to automation failures is therefore not known.

Consequently, the last study of this thesis attempts to overcome those limitations and combines primary evaluative criteria of function allocation with the secondary evaluative criterion of reliability to investigate their impacts on performance variables as well as cognitive demands in terms of attention allocation. Of special interest is the question, if interaction effects of function allocation and reliability occur that differ from findings regarding only one of the two factors.

Therefore, the last study's objective is to evaluate effects of specific combinations of function allocation and reliability on performance and cognitive demands.

1.4 Structure of the Current Research

In the following, the three studies will be presented according to the research objectives outlined in the previous chapters. The meta-analysis is described in chapter 2. The first laboratory experiment, which deals with the impact of reliability on participants' performance and cognitive demands, is presented in chapter 3. Chapter 4 introduces the experimental study evaluating differential effects of specific combinations of function allocation and reliability. Figure 3 provides a graphical overview of the three studies including research objectives, the theoretical framework, and a short description of methods.



Objective	**Theoretical Framework**	**Method**
Study I Evaluate Impacts of **Function Allocation** on Performance and Cognitive Demands	**Framework for Stages and Levels of Automation** Flow Chart Model of Automation Design: Consideration of **Primary Evaluative Criteria** **(Parasuraman, Sheridan & Wickens, 2000)**	**Meta-analysis** UV: **Function allocation** AV: Routine system performance Return-to-manual performance Workload Situation Awareness
Study II Evaluate Impacts of **Reliability** on Performance and Cognitive Demands	Flow Chart Model of Automation Design: Consideration of **Secondary Evaluative Criteria** **(Parasuraman, Sheridan & Wickens, 2000)**	**Laboratory experiment** UV: **Automation reliability** AV: Human-automation performance Concurrent task performance Visual attention allocation
Study III Evaluate Effects of Combinations of **Function Allocation and Reliability** on Performance and Cognitive Demands		**Laboratory experiment** UV: **Function allocation, automation reliability** AV: Joint human-automation performance Concurrent task performance Visual attention allocation

Figure 3. Graphical overview of the three studies conducted within the thesis

The studies constituting chapters 2 and 3 were published in peer reviewed international journals. The meta-analysis was published in Human Factors (Onnasch, Wickens, Li, &

Manzey, 2014), the first experimental study in the International Journal of Human-Computer Studies (Onnasch, Ruff, & Manzey, 2014), in which also the last experiment has been submitted for publication (under review).

A general discussion will close this work (chapter 5) containing a recapitulation of results, which are then related to the flow chart model proposed by Parasuraman et al. (2000). Furthermore, the three studies will be subject to a critical reconsideration of the applied methodical approaches. The thesis concludes with an outlook regarding future research opportunities.

1.5 References

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency." *Human Performance, Situation Awareness, and Automation: Current Research and Trends*, 54–59.

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699.

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 in Ergonomics Science*, *8*(4), 321–348.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779.

Billings, C. E. (1991). *Human-centered aircraft automation: A concept and guidelines* (technical memorandum No. 103885). NASA (Retrieved from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19910022821.pdf).

Billings, C. E. (1997). *Aviation Automation: The search for a human-centered approach*. Mahaw, NJ: Lawrence Erlbaum Associates, Inc.

Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, G., & Huff, E. M. (1976). *Aviation safety reporting system*. Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center.

Breton, R., & Bossé, É. (2003). The cognitive costs and benefits of automation. *Proceedings of the RTO HFM Symposium on "The Role of Humans in Intelligent and Automated Systems", RTO-MP-088.*

Bright, J. R. (1958). Does automation raise skill requirements? *Harvard Business Review*, *36*(4), 85–98.

Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the 34th Annual Meeting of the Human Factors & Ergonomics Society* (1524–1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Cummings, M. L., & Mitchell, P. J. (2007). Operator scheduling strategies in supervisory control of multiple UAVs. *Aerospace Science and Technology*, *11*(4), 339–348.

Dekker, S. W., & Woods, D. D. (2002). Maba-maba or abracadabra? Progress on human–automation co-ordination. *Cognition, Technology & Work*, *4*(4), 240–244.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*(3), 474–486.

Dixon, S. R., Wickens, C. D., & Chang, D. (2004). Unmanned aerial vehicle flight control: False alarms versus misses. *Proceedings of the 48th Annual Meeting of the Human Factors & Ergonomics Society*, (152–156). Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R. (1987). The application of human factors to the development of expert systems for advanced cockpits. *Proceedings of the 31st Annual Meeting of the Human Factors & Ergonomics Society* (1388–1392). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the 32nd Annual Meeting of the Human Factors & Ergonomics Society* (97–101). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64.

Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (163–181). Hillsdale, NJ: Lawrence Erlbaum Associates.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394.

Fox, J. E. (1996). The effects of information accuracy on user trust and compliance. *Conference Companion on Human Factors in Computing Systems* (35–36). New York, NY: ACM.

Galster, S. M. (2003). *An examination of complex human-machine system performance under multiple levels and stages of automation*. (No. AFRL-HE-WP-TR-2003-0149). Human Effectiveness Directorate, Air Force Research Lab, Wright-Patterson AFB OH.

Galster, S. M., & Parasuraman, R. (2004). Task dependencies in stage-based examinations of the effects of unreliable automation. *Proceeding of the Second Human Performance, Situation Awareness and Automation Conference, 2*, 23–27.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127.

Hollnagel, E., & Bye, A. (2000). Principles for modelling function allocation. *International Journal of Human-Computer Studies*, *52*(2), 253–265.

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. *Journal of Applied Psychology*, *47*(3), 161–165.

Kaber, D. B., & Endsley, M. R. (1997). Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. *Process Safety Progress*, *16*(3), 126–131.

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *10*(4), 409–430.

Layton, C., Smith, P. J., & McCoy, C. E. (1994). Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation. *Human Factors*, *36*(1), 94–119.

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

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lorenz, B., DiNocera, F., Röttger, S., & Parasuraman, R. (2002). Automated fault-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine*, *73*(9), 886–897.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*(4), 563–572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*(2), 196–204.

Milgram, P., Rastogi, A., & Grodski, J. J. (1995). Telerobotic control using augmented reality. *Proceedings of the 4th IEEE International Workshop on Robot and Human Communication*, 21–29.

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

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other. *Automation and Human Performance: Theory and Applications*, 201–220.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, *8*(1), 47–63.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5–6), 527–539.

Nof, S. Y. (2009). *Springer Handbook of Automation*. Dordrecht, New York: Springer.

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators' adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies, 72*, 772-782.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*(3), 476–488.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology*, *3*(1), 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

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

Raouf, A. (1988). Effects of automation on occupational safety & health. *Proceedings of the First International Conference on Ergonomics of Hybrid Automated Systems I*, 631–638.

Riley, V. (1989). A general model of mixed-initiative human-machine systems. *Proceedings of the 33rd Annual Meeting of the Human Factors & Ergonomics Society* (124–128). Santa Monica, CA: Human Factors and Ergonomics Society.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76–87.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. *Proceedings of the 46th Annual Meeting of the Human Factors & Ergonomics Society* (327–331). Santa Monica, CA: Human Factors and Ergonomics Society.

Salvendy, G. (1997). *Handbook of human factors and ergonomics* (second edition). New York, NY: John Wiley & Sons.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*(4), 573–583.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.) *Handbook of Human Factors and Ergonomics*, *2*, 1926–1943.

Satchell, P. (1998). *Innovation and automation*. Aldershot: Ashgate.

Sheridan, T. B. (2000). Function allocation: Algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies*, *52*(2), 203–216.

Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. New York, NY: John Wiley & Sons.

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. (Technical Report). Cambridge, MA: MIT Man-Machine Systems Laboratory.

Singh, A., Tiwari, T., & Singh, I. (2009). Effects of automation reliability and training on automation-induced complacency and perceived mental workload. *Journal of the Indian Academy of Applied Psycholgy*, *35*, 9–22.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Wickens, C. D. (2000). *Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness*. (Tech. Rep. ARL-00-10/NASA-00-2). Savoy: University of Illinois, Aviation Research Lab.

Wickens, C. D., Dixon, S. R., Goh, J., Hammer, B. (2005). Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. *Proceedings of the 13th International Symposium on Aviation Psychology* (919–923). Columbus, OH: Association of Aviation Psychology.

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.

Wickens, C. D., Hollands, J., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology & human performance* (4 edition.). Boston: Pearson.

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 Academy Press.

Wiener, E. L. (1981). Complacency: Is the term useful for air safety. *Proceedings of the 26th Corporate Aviation Safety Seminar* (116–125). Denver, CO: Flight Safety Foundation.

Wiener, E. L. (1988). Cockpit automation. In E. L. Wiener & D. C. Nagel (Eds.), *Human factors in aviation* (433–461). San Diego, CA: Academic Press.

Wiener, E. L. (1989). Human factors of advanced technology (“glass cockpit”) transport aircraft (Tech. Report 177528). Moffett Field, CA: NASA Ames Research Center.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*(10), 995–1011.

2. Study I

Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis

Linda Onnasch[1], Christopher D. Wickens[2], Huiyang Li[3], Dietrich Manzey[1]

[1]Berlin Institute of Technology, Berlin, Germany

[2]Alion Science and Technology, McLean, VA

[3]University of Michigan, Ann Arbor, MI

Reference

Onnasch, L., Wickens, C. D., Li, H. & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors*, *56*(3), 476-488.

2.1 Abstract

Objective: We investigated how automation-induced human performance consequences depended on the degree of automation (DOA). **Background:** Function allocation between human and automation can be represented in terms of the stages & levels taxonomy proposed by Parasuraman, Sheridan & Wickens (2000). Higher DOAs are achieved both by later stages and higher levels within stages. **Method:** A meta-analysis based on data of 18 experiments examines the mediating effects of DOA on routine system performance, performance when the automation fails, workload, and situation awareness (SA). The effects of DOA on these measures are summarized by level of statistical significance. **Results:** We found: (1) a clear automation benefit for routine system performance with increasing DOA, (2) a similar but weaker pattern for workload when automation functioned properly, (3) a negative impact of higher DOA on failure system performance and SA. Most interesting was the finding that negative consequences of automation seem to be most likely when DOA moved across a critical boundary which was identified between automation supporting information analysis and automation supporting action selection. **Conclusions:** Results support the proposed cost-benefit trade-off with regard to DOA. It seems that routine performance and workload on the one hand, and the potential loss of SA and manual skills on the other hand, directly trade-off and that appropriate function allocation can only serve one of the two aspects. **Application:** Findings contribute to the body of research on adequate function allocation by providing an overall picture through quantitatively combining data from a variety of studies across varying domains.

Keywords: degree of automation, operator performance, workload, situation awareness, human-automation interaction, function allocation

2.2 Introduction

It has been long known that automation can both hurt and benefit human performance (e.g., Bainbridge, 1983; Wiener & Curry, 1980; Sheridan, 2002; Wickens, Mavor, Parasuraman & McGee, 1998; Ephrath & Young, 1981; Kessel & Wickens, 1982; Wickens & Kessel, 1979, 1981; Rasmussen & Rouse, 1981). This cost-benefit trade-off is particularly prominent when automation is imperfectly reliable. Automation infrequently fails, either due to hardware or software failures, or it fails to achieve desired outcomes simply because a functionality is used in circumstances for which it was not intended. For fielded automation, it is almost always the case that routine or "non-failure" performance substantially exceeds unaided human performance and/or the automation assistance lowers workload. If it did not, the system would not be fielded or considered useful.

However, on those infrequent occasions when automation does fail, the effects on joint human-machine system performance may be catastrophic. These catastrophic effects may result from human's reduced monitoring of highly reliable automation at the time it fails, trusting it too much (Parasuraman & Riley, 1997), and losing situation awareness (Endsley & Kiris, 1995). This is sometimes described as a form of complacency (Parasuraman, Molloy & Singh, 1993) or an automation-induced decision bias (Mosier & Skitka, 1996). Indeed, operators occasionally over-rely on automation and exhibit complacency because the highly (but not perfectly) reliable automation functioned properly for an extended period of time prior to this first failure (Parasuraman et al.1993; Parasuraman & Manzey, 2010; Yeh, Merlo, Wickens & Brandenburg, 2003). Going beyond the issues of highly reliable automation, Endsley and Kiris (1995) and Miller and Parasuraman (2007) have pointed out, that also the "competence" of the automation must be considered. The more support an automated system provides, the higher the risk of adverse effects on human performance (e.g., complacency, loss of situation awareness, skill degradation), and the greater the likelihood of catastrophic consequences when it fails. This trade-off, in which *more automation yields better human-system performance when all is well but induces increased dependence which may produce more problematic performance when things fail*, will be of critical importance to this review of the performance effects of different degrees of automation. We might refer to this conventional wisdom about automation as the "lumber jack effect"; as applied to trees in the forest: "the higher they are, the farther they fall". Importantly, the choice of whether or not, and to what degree, to automate a particular function should involve a trade-off between the benefits of reliable automation and the expected costs (true costs x probability of failure) of automation failures (Sheridan & Parasuraman, 2000).

The "routine-failure" trade-off is complicated by the fact that "automation" is not an all-or-none concept, as it was often assumed to be in the classic human-machine task allocation analyses (e.g., the "Fitts List"; for a critique of those analyses, see Dekker &Woods, 2002; Parasuraman, Sheridan, & Wickens, 2008). Instead, one can think of varying **levels of automation** as first put forth by Sheridan & Verplank, (1978; see also Endsley & Kiris, 1995). This continuum can be jointly defined by the amount of automation autonomy and responsibility (highest at the highest level) and the amount of human physical and cognitive activity (highest at the lowest level). For example, at the highest level, the automation can perform a decision task completely autonomously; at a lower level, it can choose (and possibly execute) an option unless the human vetoes; and at an even lower level, it may simply offer the human a selection of options.

More recently, Kaber & Endsley (1997), and Wickens et al, (1998), put forth the idea that automation could also be categorized according to the **stage of information processing** that it accomplished. Elaborating upon this concept, Parasuraman, Sheridan, and Wickens (2000), and Wickens et al. (1998) proposed a concept in which automation could **filter** information from the environment (stage 1: information acquisition), i**ntegrate** this information, as when forming an assessment based on several sources of information (stage 2: information analysis), choose or **decide** upon an action based on the assessment (stage 3: decision and action selection) **and implement** the action via a typically manual activity (stage 4: action implementation). Within each stage, varying levels could be defined. For example, as described above, Sheridan & Verplank (1978) define multiple levels at stage 3. In so doing, automation can be said to offload, assist or replace human performance at corresponding stages of human information processing (e.g., automation filtering at stage 1, can assist human selective attention).

As an example, health care automation may (1) alert (call attention to) abnormal patient symptoms, (2) integrate these symptoms to form an intelligent diagnosis of the patient condition, (3) recommend a treatment or course of action based upon the diagnosis, and (4) carry out the action as, for example, with an automated infusion pump. In applying this taxonomy, where any given stage can function at various levels, it is important to note the quasi-independence of levels across the various stages. Thus, for example, a totally automated diagnosis, may be followed by a fully manual (physician chosen) course of action; just as a fully manual diagnosis may trigger an automated choice of treatment.

Considering that "more automation" can be represented both by higher levels within a stage, and, typically, later stages (which, in literature, are typically preceded by automation at earlier stages), we assume, in the analysis below, that these two dimensions (higher levels and later stages) increase the *degree of automation* (DOA; e.g., Manzey, Reichenbach & Onnasch, 2012)**.** More specifically, it is assumed that differences between automated (support) systems representing automation of different stages and levels can be described on an ordinal scale reflecting the amount of automated support which is provided. The main assumption underlying this concept as we define it, asserts that assessment of "more vs. less automation" can be based on dominance relationships, as long as the following three postulates are agreed upon. That is, all other factors held equivalent (1) a higher **level** of automation constitutes "more automation", (2) a later **stage** of automation constitutes "more automation", and (3) as a consequence, a combination of higher levels and a greater **number** of stages at which automation is implemented constitutes "more automation". As will be shown below, applying this reasoning enables an unambiguous rank ordering of automated systems which have been analyzed and compared in different studies and domains. It furthers seems to reflect the implicit or explicit assumptions which researchers in the field whose data are employed in the current analysis usually apply when comparing their systems in terms of some concept of "more or less automation".

We illustrate this within Figure 1 which, for simplicity, presents examples of the four stage model of Parasuraman et al. (2000) with only three levels per stage. Each of the four cases contains two automation systems, A and B, which are compared within one experiment. The automation characteristic of each of these systems is characterized by a profile of levels across the stages. The first three cases also represent the three postulates above. Case #1 ("Pure Levels"): different levels within a stage. Case #2 ("Pure Stages"): different stages at the same level. Case #3 ("Aggregation"): an earlier stage and lower level vs. a later stage and higher level. Case 4 ("Confound"): an earlier stage and higher level vs a later stage and lower level (i.e., a "tradeoff" between stages and levels).

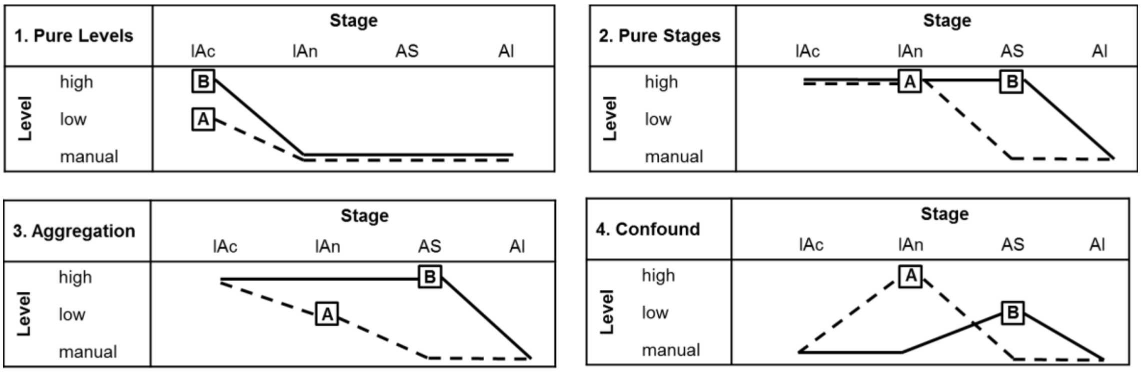


Figure 1. Four cases comparing DOA across Stages, i.e. Information Acquisition (IAc), Information Analysis (IAn), Action Selection (AS) and Action Implementation (AI), and Levels, i.e. high, low and manual. Two systems compared are represented by dashed (System A) and solid lines (System B). For cases 1 – 3, System B always represents "more automation" in a distinct way (e.g., 1. by higher levels, or 2. by higher stages). Case 4 represents a confound where "what is more automation" cannot be defined.

We argue that, to the extent that the three postulates above are agreed upon, all of the comparisons 1-3 clearly represent contrasts between systems with more (system B) vs. less (system A) automation, as defined on an ordinal scale. These relationships characterize all of the studies we have reviewed in which authors have invoked a phrase like “more automation”.

Case 4 is an important exception. Here, there is a trade-off between later stages and higher levels. It is impossible to assess a relative DOA unless both stages and levels are expressed on an interval or ratio scale, and we have no confidence that this has or even can be done. But none of the studies analyzed below involved such a comparison.

Thus, in our analysis, DOA is a useful ordinal metric explicitly available and used to compare two or more systems (or experimental conditions), specifically for the purpose of examining the trade-offs inherent in the lumberjack analogy.

Within this DOA concept, the discrete trade-off described above (i.e., automation supports better performance in routine situations but is problematic when automation breaks down) can be expressed as a more continuous trade-off, as illustrated in Figure 2. The two primary performance functions in this figure (heavy lines) indicate that, as the degree of automation increases, routine performance will improve but performance under failure will decline. This relationship is expressed intuitively by the lumberjack analogy. Prior research has found that the lumberjack analogy appears to apply to the continuum of automation reliability of alerting systems (Wickens & Dixon, 2007). We examine here the extent to which this may also apply to DOA. Furthermore, our interest lies in whether DOA also has a systematic impact on workload and situation awareness (Endsley & Kiris, 1995). Indeed, as discussed below, to the extent that loss of situation awareness may be due to both - an increase in automation reliability, and an increased degree of automation- it is plausible to assume that the lumberjack analogy may apply to the latter case (see also Wickens, 2008a).

Thus, Figure 2 also depicts the hypothetical trade-off between the two secondary variables, workload and loss of situation awareness (the two lighter lines). With a higher degree of automation, the workload imposed by the automated task is progressively reduced, almost by definition, since if the automation is doing more cognitive/ physical work, the human is doing less. This holds at least if the automation is properly designed and does not provide new effortful challenges and tasks related to its engagement and monitoring (e.g., Kirlik, 1993; Wiener, 1988). If this is granted, the automation enables the human to allocate more attention to other concurrent tasks (Wickens, 2008b); but if the human does so (i.e., exploits the lower workload to enhance overall productivity), the resulting reduction of attention to the tasks served by automation could have consequences expressed in the loss of situation awareness (LSA); that is, loss of awareness of the state of the system supported by automation (e.g., lack of altitude awareness in the autopilot-controlled cockpit), or the state of the automation itself (i.e., poor mode awareness of the flight management system; e.g.,Sarter, 2008).

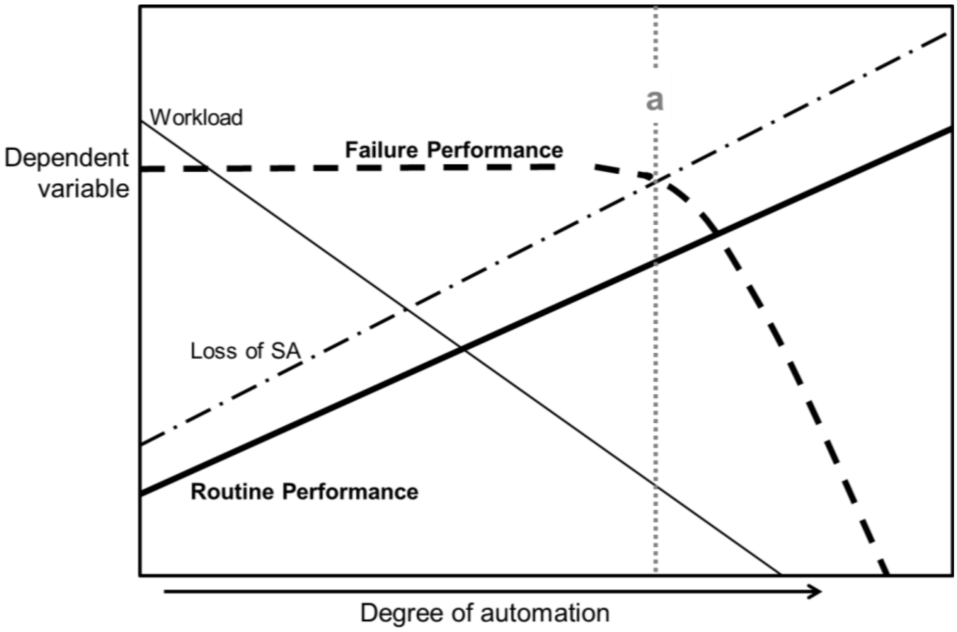


Figure 2. Trade-off of variables, with degree of automation

Even though there is broad consensus in the understanding of the concept of situation awareness (SA) as it has been defined by Endsley's three levels model (1988), the operational definitions used to assess SA in different studies are considerably diverse. In the context of the present research we consider both direct as well as indirect measures as indicators of SA. Direct indicators are derived from conventional methods to assess SA, like the Situation Awareness Global Assessment Technique (SAGAT, Endsley, 2000). Indirect measures of (a loss of) SA include any performance consequences in interaction with automation that point to a lack of information sampling, a lack of understanding or a lack of correctly anticipating the behavior of the automation (e.g.,errors of omission or commission, Mosier & Skitka, 1996).

The hypothetical trade-offs depicted in Figure 2 are critical for task allocation because these trade-offs may not be linear, and in some cases, a "flat" function may allow strong recommendations for the optimal task allocation (Wickens, 2008). For example, if the costs of imperfect automation (mediated by LSA) remain flat up to a high degree of automation (as shown in Figure 2), then the recommended degree of automation would be at point (a) in the Figure: maximum routine performance and lowest workload, without sacrificing failure performance.

Earlier research contrasting human performance with and without automation support only have focused on what has been referred to as "out-of the-loop unfamiliarity" effects without varying the levels or stages of automation (e.g.,Crossman, 1974; Eprath & Young, 1981; Kessel & Wickens, 1982; Wickens & Kessel, 1979; 1980, 1981). These studies provide evidence for automation-induced performance consequences but do not allow for any conclusion about the relationship to different degrees of automation. The latter issue attracted little research until the early 1990s (see for early examples, e.g., Crocoll & Coury, 1990; Layton, Smith & McGoy, 1994). Yet, since then at least a limited number of studies have become available that have collected empirical data on effects of two or more different DOAs on workload and/or SA (e.g., Endsley & Kiris, 1995; Kaber, Onal & Endsley, 2000; Lorenz, Di Nocera, Röttger & Parasuraman, 2002a; Sarter & Schroeder, 2001). The pattern of results of these single studies provides a somewhat mixed picture. Whereas some studies support the existence of the above trade-off as defined by better routine performance but worse performance when automation fails (e.g.,Sarter & Schroeder, 2001) others do not find this effect (Lorenz et al., 2002a) and still others suggest that medium levels of automation provide the best choice in terms of maintaining SA and return-to-manual performance (Endsley & Kiris, 1995) or provide an even more complex pattern of effects (Endsley & Kaber, 1999). However, due to differences in DOA levels considered, and a generally limited statistical power, the effects of single studies are inconclusive. A more valid overall picture might be revealed by quantitatively combining data from a variety of studies across varying domains (e.g., process control, aviation), an approach analogous to a classic meta-analysis (Rosenthal 1991; Fadden, Ververs, & Wickens, 1998; Horrey & Wickens, 2006; Wickens, Hutchinson, Carolan & Cumming, 2013). The purpose of the current investigation is to provide such meta-analysis by (a) aggregating data from studies that compared different degrees of automation, (b) examining the extent to which they show the postulated trade-off between normal operations and failure conditions as the degree of automation (DOA) was manipulated and (c) if possible, by identifying factors that may mitigate or moderate this trade-off.

2.3 Methods

In a first step we looked for relevant studies to be included in this analysis. Sources used for this purpose included databank searches (e.g.,PsychInfo), analyses of tables of content of relevant journals (e.g.,Human Factors, Ergonomics, International Journal of

Human-Computer Interaction), conference proceedings for the years 1990ff , and direct contact of colleagues in order to identify relevant technical reports or other examples of references that were not available through publishers. Only studies which compared at least two different degrees of automation defined by the postulates above, e.g., either by varying the stage of automation, the number of stages or by varying the level of automation within a stage, with respect to at least one relevant performance measure were included. Consequently, a total of 18 studies was identified and integrated in the analysis (see Table 1).

The second step included a proper quantification of the independent variable (i.e. DOA) and dependent variables (i.e. performance, workload, and SA data) as a basis for our meta-analysis approach. For each single study, the DOAs analyzed were converted into rank data with an increasing rank (beginning by rank= 1) reflecting an increasing DOA via either stages or levels corresponding to the logic described above. We note that none of the studies contrasted conditions with higher level/ earlier stage with lower level/ later stage (or vice versa) which would not be easy to rank due to lacking unambiguous a priori criteria for cross-stage comparisons of levels. Manual performance conditions were always assigned rank = 0. Rankings were provided by one of the authors (LO) and double-checked by two of the co-authors (CW, DM).

To bring the variety of dependent measures and definitions used in the studies to a comparable level we defined ‘meta-variables’ which were broad enough to group the data while still representing a clear definition of the concept in question (e.g.,situation awareness). As our main focus of the present study was on performance costs and benefits of automation support we differentiated between primary task performance when the automation functioned properly (meta-variable *routine primary task performance*, reflecting joint performance of operator and system together) and performance when there was a complete automation breakdown, i.e. when operators had to resume the automated task and perform it manually again after some time of reliable automation support (meta-variable *return-to-manual primary task performance*). The meta-variable *routine primary task performance*, for example, could be realized within the single studies as fault identification time in a monitoring task (e.g., Lorenz, DiNocera, Röttger & Parasuraman, 2002b), the decision accuracy in interaction with an automated decision aid (e.g., Rovira, McGarry & Parasuraman, 2007) or the out-of-target error when the main task was to maintain certain values in a dynamic task (e.g., Manzey et al., 2012). Nevertheless, all these measures represented operators' performance when working together with a reliable automation support and were therefore subsumed under the same meta-variable. For defining the meta-variable *return-to-manual primary task performance* the same measures as for routine performance were considered for a given study but for a situation, where the operator needed to perform the primary task manually after a complete automation breakdown of the automation.

Workload measures were assessed in two different ways: As a performance variable we defined the meta-variable *secondary task performance* (if the study used a multi-task environment) again for routine and return-to-manual performance, respectively. A second meta-variable was operators' *subjective workload*, typically quantified by the NASA-TLX (Hart & Staveland, 1988) as used for example by Endsley & Kaber (1999).

The meta-variable *situation awareness* merged any direct and indirect indicators that pointed to a loss of situation awareness when working together with automation. As direct indicators we considered the outcome of techniques that are designed to directly ask for SA like the SAGAT (Endsley, 1988; 2000) or questionnaires as SART (Taylor, 1990). As indirect evidence for a possible loss of SA we considered all sorts of operators' errors which might be attributed to a loss of SA due to an overtrust in automation or a lack of proper understanding. Such errors can include, e.g., mode errors (Sarter, 2008), or errors of omission or commission (Mosier & Skitka, 1996), i.e. errors where operators failed to respond to a critical situation if the automation failed to alert them properly or where operators followed wrong advice of automation without detecting this failure. When participants committed these kinds of error we interpreted this as evidence for deficient situation awareness as they did not realize that the automation had made a mistake.

Departing from the classic meta-analysis approach we assigned rankings for every meta-variable within a single study according to significant effects found with regard to DOA (a priori, a posteriori). This was done as effect sizes (e.g., Hedges g) were only rarely reported in the original studies and therefore could not be used for analysis, without eliminating many studies from consideration. Furthermore, any other estimates of effect sizes based on the F ratios for multiple conditions reported in the studies would not be able to capture the ordinal aspect of data which is of particular relevance for our question. Although unconventional, this approach of data aggregation is in line with the basic idea of meta-analysis (e.g., Rosenthal, 1991) where no particular statistical method is defined for this “analysis of analyses”. It is also in line with other authors who also departed from the classic approach for similar reasons (e.g., Wickens & Dixon, 2007; Wickens, Hooey, Gore, Sebok, & Koenicke, 2009; Wickens, Hutchins, Carolan, & Cumming, 2013; Hutchins, Wickens, Carolan, & Cumming, 2013).

Different rankings were assigned when there was a significant effect between two DOA conditions ($p < .05$). In case of non-significant effects between different degrees of automation we assigned tied ranks. For example, in case a study comparing the impact of three different DOAs on *routine primary task performance* revealed all pairwise comparisons between the DOA conditions as significant, the condition showing the worst performance was assigned rank 1, the condition with the second-best performance rank 2 and rank 3 was assigned to the condition with best performance. However, when only one condition differed in terms of superior performance compared to the other two conditions, the best condition was assigned rank 3 and the other two conditions were assigned tied ranks, in this case rank 1.5.

When a meta-variable was measured by more than one dependent variable within a study (e.g., error of omission and SAGAT for situation awareness) the rankings of the single variables were integrated into one 'overall-ranking'. With this approach we were able to integrate data from various studies assessed in numerous ways to examine the trade-off when automation degree increased and to identify trends on a descriptive level.

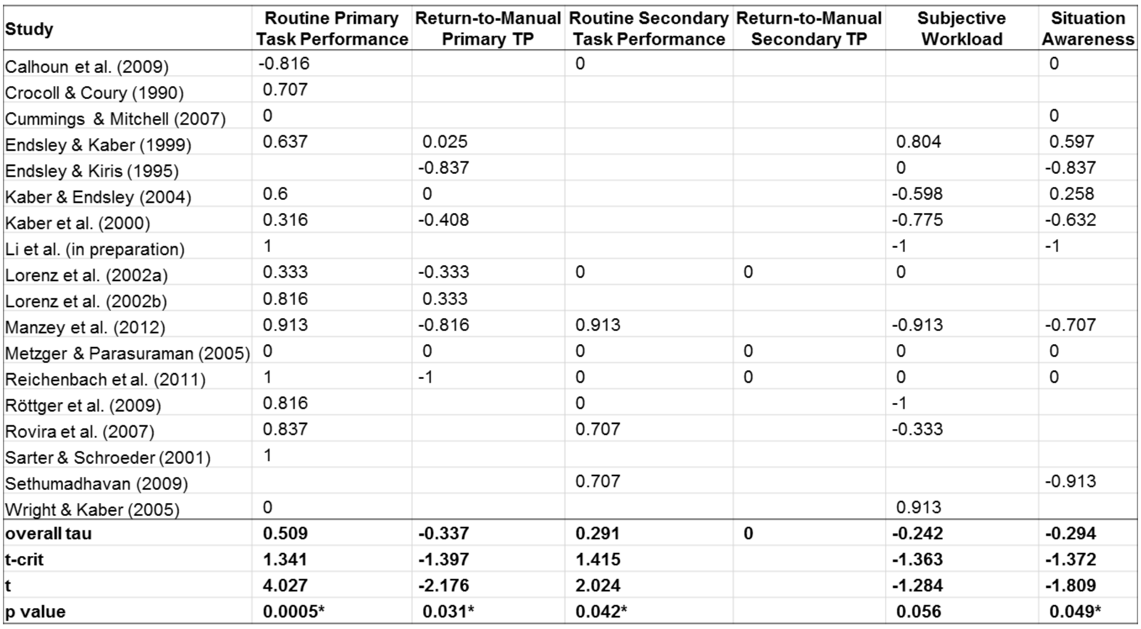
In a third step we described the relationship between the DOAs and the different meta-variables by computing Kendall's *tau* (correlation between rank orderings) to see if the DOA had an impact on a certain class of meta-variable. Kendall's *tau* was used as an alternative analysis to product moment correlations as we only had rank orderings as data bases. With this analysis it was possible to determine and test for a monotonic relation between two dependent variables (e.g., DOA and workload). Furthermore, Kendall's *tau* does not make the implicit assumption of equidistance between different rankings which would not have been the case for our data.

To further abstract the gained results we computed an overall Kendall's *tau* for every meta-variable across studies and tested with one-tailed t-tests if this correlation was different from zero in the hypothesized direction. In doing so, we defined each Kendall's *tau*, computed for every study, as a certain manifestation of the variable in question (e.g., *routine primary task performance*). With this last step, we could also examine various instances of the trade-off: For example, do the routine and failure aspects of performance trade off? How strongly is decreased failure response coupled with LSA? Do workload and LSA trade off?

2.4 Results

Table 1 shows the correlations of DOA on the six meta-variables for the single studies (Kendall's *tau*) and the computed overall Kendall's *tau* for every meta-variable including statistics of one-tailed t-tests.

Table 1. Kendall's tau for the single studies on the six meta-variables with resulting overall Kendall's tau and statistics of one-tailed t-tests.



Study	Routine Primary Task Performance	Return-to-Manual Primary TP	Routine Secondary Task Performance	Return-to-Manual Secondary TP	Subjective Workload	Situation Awareness
Calhoun et al. (2009)	-0.816		0			0
Crocoll & Coury (1990)	0.707					
Cummings & Mitchell (2007)	0					0
Endsley & Kaber (1999)	0.637	0.025			0.804	0.597
Endsley & Kiris (1995)		-0.837			0	-0.837
Kaber & Endsley (2004)	0.6	0			-0.598	0.258
Kaber et al. (2000)	0.316	-0.408			-0.775	-0.632
Li et al. (in preparation)	1				-1	-1
Lorenz et al. (2002a)	0.333	-0.333	0	0	0	
Lorenz et al. (2002b)	0.816	0.333				
Manzey et al. (2012)	0.913	-0.816	0.913		-0.913	-0.707
Metzger & Parasuraman (2005)	0	0	0	0	0	0
Reichenbach et al. (2011)	1	-1	0	0	0	0
Röttger et al. (2009)	0.816		0		-1	
Rovira et al. (2007)	0.837		0.707		-0.333	
Sarter & Schroeder (2001)	1					
Sethumadhavan (2009)			0.707			-0.913
Wright & Kaber (2005)	0				0.913	
overall tau	**0.509**	**-0.337**	**0.291**	**0**	**-0.242**	**-0.294**
t-crit	**1.341**	**-1.397**	**1.415**		**-1.363**	**-1.372**
t	**4.027**	**-2.176**	**2.024**		**-1.284**	**-1.809**
p value	**0.0005***	**0.031***	**0.042***		**0.056**	**0.049***

Primary Task Performance

16 studies provided data for the meta-variable *routine primary task performance*. In terms of Kendall's *tau*, a vast majority of these studies indicated a strong positive correlation of DOA and routine performance. This is in accordance with the anticipated benefit of automation support with increasing DOA when automation functioned properly. Data of one study only resulted in a negative correlation and additional three studies revealed no evidence for a relation of DOA and *routine primary task performance*. Looking at the amount of studies with positive *tau*s and the strength of these correlations supports the hypothesized benefit of automation support with increasing DOA. This interpretation is also backed up by a significant overall rank correlation across studies of *tau*=0.51, $p < .001$.

For an assessment of the impact of DOA on *return-to-manual primary task performance* data of nine studies were available. Five of these studies reported effects that resulted in a negative Kendall's *tau* whereas only three others showed no evidence for the hypothesized negative impact of DOA when participants had to resume the formerly automated task because of an automation breakdown. This general trend was reinforced by a negative overall Kendall's *tau* averaged across the nine studies that was significantly different from zero, *tau*=-0.34, *p* = .03.

Taken together results for primary task performance support the hypothesized lumberjack effect as the routine and failure aspects of performance trade off with increasing automation complexity. Further in line with the hypothesized trade-off (Figure 2) is the fact that eight out of the nine studies that assessed both aspects of performance (routine and return-to-manual) showed a higher (more positive) correlation of DOA with routine than with failure performance, and the single exception (Metzger and Parasuraman, 2005) showed a zero correlation in both cases.

Workload

Workload was evaluated on a performance level and on a subjective level. Eight studies provided data for the meta-variable *routine secondary task performance*. Three out of these eight studies revealed a strong positive correlation of DOA and performance, i.e. operators showed better results when supported by higher degrees of reliable automation. This was also supported by a significant overall Kendall's *tau* = +.03; *p* = .04. However, this interpretation is challenged by the fact that five out of the eight studies revealed no connection between DOA and performance in terms of zero correlations. Therefore, results have to be interpreted with caution.

In contrast to primary task performance, secondary task performance did not seem to be affected by surprising automation breakdowns, as there was no evidence for an impact of DOA on *return-to-manual secondary task performance*. However, only three studies reported data for this variable so that the explanatory power of this result is rather low.

Concerning the impact of DOA on *subjective workload*, the 12 studies which reported data for this meta-variable provided a quite complex pattern of results. Two out of these studies (Endsley & Kaber, 1999; Wright & Kaber, 2005) reported data that revealed strong positive relations between DOA and *subjective workload* (*tau* = +.08, *tau* = +.913). In contrast, six studies provided a reversed pattern with strong negative Kendall's *tau*s, and the remaining four studies showed no correlation at all.

However, because the majority of data provided negative correlations, overall Kendall's *tau* also showed a negative, albeit weak trend (*tau* = -.24, p=.05) that supports the often stated argument that higher degrees of automation reduce operators' workload. Nevertheless, because of the different results of the single studies further research is needed to assure the proposed interpretation.

Situation Awareness

We hypothesized that one of the costs concerned with higher degrees of automation would be associated with a loss of *situation awareness*. 11 studies reported data for this meta-variable. Whereas five studies, like Endsley & Kiris (1995) or Manzey et al. (2012), did report a potential loss of SA with increasing automation, four other studies did not find an impact of DOA (Calhoun et al., 2009; Cummings & Mitchell, 2007; Reichenbach et al., 2011; Metzger & Parasuraman, 2005), and the data of the two remaining studies even resulted in positive correlations of DOA on *situation awareness*. Due to this, the hypothesized negative trend was not as strong as expected, with an overall Kendall's *tau* =-.29, but still significantly different from zero (p=.04).

Taking a closer look on the single studies, it is striking that the two studies with the highest lumberjack trade-off (*routine vs. return-to-manual primary task performance*; Manzey et al., 2012; Kaber, Onal & Endsley, 2000) were also two of the four studies that yielded comparatively strong negative correlations between DOA and *situation awareness* (values of -0.71 and -0.63, respectively). This is in accordance with the assumption that higher DOAs increase the risk of out-of-the-loop unfamiliarity issues reflected in a loss of SA as well as with negative performance consequences in case an operator unexpectedly needs to resume manual control of an automated task (Endsley & Kiris, 1995). Similarly, the strongest negative correlation between DOA on SA was found in the study conducted by Li et al. (in preparation) which at the same time showed the greatest automation benefits *for routine primary task performance* as well as the greatest decrease in *subjective workload* of all studies. Therefore, it seems that *routine primary task performance* and workload on the one hand, and the potential loss of SA, on the other hand, directly trade off and that appropriate function allocation can only serve one of the two aspects.

Moderating Factors

In a next step we tried to identify possible factors that might moderate potential trade-offs between the different measures. We looked for aspects that some studies had in common, especially those that strongly supported the trade-off hypothesis, but also differentiated them from other studies. As one such variable, we focused on the critical distinction between automation that supported situation assessment by providing automated information acquisition and analysis (stages 1 or 2) versus that which supported the selection and execution of action (stages 3 and 4). This distinction of assessment versus action is a ubiquitous one that underlies many facets of human performance (Wickens, Hollands, Banbury & Parasuraman, 2012). For this analysis only those studies were included which varied the DOA across the assumed critical boundary from information analysis support to action selection support (i.e. Crocoll & Coury, 1990; Cummings & Mitchell, 2007; Reichenbach, Onnasch & Manzey, 2011; Manzey et al., 2012; Rovira, McGarry & Parasuraman, 2007; Sarter & Schroeder, 2001). Examining these studies exclusively, we found that when DOA was varied across this boundary, the pattern of the lumberjack analogy trade-off was substantially amplified. Calculated for these six studies separately, the overall Kendall's *tau* correlation of DOA with routine performance was +0.68, higher than the overall correlation of 0.51 for all studies (see Table 1); and the overall correlation with return-to-manual performance was -0.90, much more negative than the overall correlation of all studies, a value of -0.34.

We also examined how other variables such as prior experience with failures, or subject experience might have modulated the trade-off. Concerning the training participants received, we looked especially for the possible impact of 'first failure automation training' which has been found to impact human performance with automation (Bahner, Hueper & Manzey, 2008). Yet, none of the integrated studies applied to such kind of training. The only systematic differences in training were related to practice time before the experiment started. However, this difference is hard to evaluate since training time usually depends on the complexity of the experimental task.

In all but four studies students served as participants. In one study participants were recruited from Air Force personnel but still were novices for the experimental task (Calhoun, Draper & Ruff, 2009). Three studies were conducted with experts as the experimental simulation was very realistic (Control of UAVs, ATC, Pilots). Nevertheless, these differences did not moderate the reported lumberjack trade-off effects of DOA.

Also, most studies used multi-task settings but differed in the number of concurrent tasks (2 or 3). Yet, four studies represented single-task studies (Crocoll & Coury, 1990; Endsley & Kiris, 1995; Endsley & Kaber, 1999; Kaber et al., 2000). One could assume that this differentiation might be important, especially for variables like workload or SA. However, the amount of secondary tasks did not seem to make a difference.

Another variable that was examined in detail was the nature of the display of the automated process. The rationale for this focus was twofold. (1) The emerging literature that clear, intuitive or "ecological" displays of the state of automated processes can support a proper response to automation failures (Bennett & Flach, 2011; Burns, Skraaning, Jamieson, Lau, Kwok, Welch et al., 2008; Seppelt & Lee, 2007). (2) The linkage between displays and SA support on the one hand, and our finding which suggests that LSA might be related to return-to-manual performance issues. While this relation too did not emerge from our post-hoc analysis of the data, we are certainly reluctant to conclude that effective displays do not support off-nominal response via the mediating role of SA, because of the relatively low power of our assessment.

2.5 Discussion

Overall, the results fairly conclusively confirm the lumberjack hypothesis with regard to the degree of automation. “Conventional wisdom” has now been transformed into “statistical wisdom”. Thus, the pattern underlying the degree of automation confirms the general pattern that had previously been observed regarding the presence or absence of automation. Automation helps when all goes well, but leaving the user out-of-the-loop can be problematic because it leads to considerable performance impairment if the automation suddenly fails. And this risk appears to increase with increasing DOA. The data presented in Table 1 further suggest that this effect is linked to raised issues of LSA with increasing DOA. However, due to a lack of statistical power this latter conclusion needs to be treated with caution.

The most promising account is suggested by the final post hoc analysis reported above. When DOA moves across the critical boundary from information acquisition and information analysis to action selection, the latter alleviating the human from some or all aspects of choosing an action, then the human is much more vulnerable to automation “failures”. Actively choosing actions manually (the generation effect, Slamecka & Graf, 1978) supports SA in a way that supports the manual performance in case of automation breakdown. When that choice is removed, the automation failure response suffers. Thus, the distinction between situation assessment and action support is critically important in automation, just as the simple dichotomy is in other aspects of human factors and cognitive engineering, such as cognitive task analysis (Hoffman, Crandall & Shadbolt, 1998), predicting multi-task performance (Wickens, 2008), and predicting transfer of training (Osgood, 1949).

This finding also qualifies and specifies earlier claims that medium levels of automation would represent an optimum choice with respect to primary performance improvements and workload reductions by, at the same time, reducing unwanted performance consequences in terms of LSA and difficulties of return-to-manual performance

(Endsley & Kiris, 1995). The direct trade-off between DOA-related consequences on primary task performance and return-to-manual performance, respectively, suggest that there is no clear optimum of automation support. Each step of increase of DOA seems to be associated with an increase of the risk of return-to-manual performance decrements, meaning that there is no specific DOA below which automation-induced performance benefits can be increased without any performance costs. This renders doubts in any simple design-recipes like “medium DOAs are best”. However, the strength of the trade-off is important particularly if the border between information and action support is crossed. That is, the general recommendation of preferring “medium levels of automation” where the human is kept somehow “in-the-loop” can now be turned into a more specific one: if return-to-manual performance issues are of serious concern, human operators should be kept involved at least to some extent in decision and action selection as well as action implementation. Although, even if in this case risks of return-to-manual might not be fully excluded they can probably be kept on a comparably low level.

One limitation of the present research is the comparably low number of studies available for this analysis and the need to just consider rank data with respect to scaling DOA and performance effects. Our approach of using rank orders based on dominance orderings of three features, i.e. stages, levels, and number of stages neither allowed for quantifying the DOA on a ratio or interval scale, nor for resolving trade-offs between stages and levels. This made it difficult to yield clear statistical conclusions for all of the findings and limits the conclusiveness of results with respect to the formal characteristics of the observed trade-offs (to what extent are they linear?). However, based on the limited current knowledge and available data, the rank order approach applied to represent DOA seemed to be the only way to yield a quantitative input for our meta-analysis. Clearly, much more psychophysical and controlled experimental research is needed before more distinct metric DOA scales may be developed. A second limitation is the possibility that we might have underestimated the trends within any particular study, by the relatively course dichotomous “grain size” by which effects were coded (significant vs. non-significant). In doing so, we collapsed across quantitative measures of the size of an effect that might have added precision to the coding. Taking these limitations in mind, the overall pattern of raw effects and statistical results provides a first quantitative summary of the state of knowledge about performance consequences of stages and levels of automation which, together with the remaining questions concerning possible moderating factors, certainly offers an invitation for future research.

2.6 Acknowledgement

This research was sponsored by NASA-Johnson Space Center Grant #NNX09AM81G. Barbara Wolford was the scientific-technical monitor. The authors wish to thank Angelia Sebok and Professor Nadine Sarter for their support in managing and monitoring the research contract under which this review was carried out.

2.7 Key Points

- Increasing DOA supports routine system performance and workload
- Increasing DOA negatively impacts failure system performance and SA
- Negative consequences of automation most likely when DOA moves from stage 2 to stage 3 automation

2.8 References (including citations in Table 1)

Bahner, E. J., Huper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and training experience. *International Journal of Human-Computer Studies, 66*, 688–699.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19,* 775-779.

Bennett, K. B., & Flach, J. M. (2011). *Display and interface design: Subtle science, exact art.* Boca Raton, FL: CRC Press.

Burns, C. M., Skraaning, G., Jamieson, G., Lau, N., Kwok, J., Welch, R. et al., (2008). Evaluation of ecological interface design for nuclear process control: situation awareness effects. *Human Factors, 50,* 663–698.

Calhoun, G. L., Draper, M. K., & Ruff, H. A. (2009). Effect of level of automation on unmanned aerial vehicle routing task. In *Proceedings Human Factors and Ergonomics Society 53rd Annual Meeting* (197-201). Santa Monica, CA: Human Factors and Ergonomics Society.

Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the Human Factors Society 34th Annual Meeting* (1524-1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Crossman, E. R. F. W. (1974). Automation and skills. In E. Edwards & F. Lees (Eds.), *The human operator in process control* (1-24). London: Taylor & Francis.

Cummings, M. L., & Mitchell, P. J. (2007). Operator scheduling strategies in supervisory control of multiple UAVs. *Aerospace Science and Technology*, *11*, 339-348.

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human-automation co-ordination. *Cognition, Technology, and Work, 4,* 240-244.

Endsley, M. R. (2000). Direct measurement of SA: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness: Analysis and measurement* (147-173). Mahwah: Lawrence Erlbaum.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the National Aerospace and Electronics Conference (NAECON)*, 789–795. New York: IEEE.

Endsley, M. R. & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*, 462-492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*, 387-394.

Ephrath, A. R. & Young, L. R. (1981). Monitoring vs. man-in-the-loop detection of aircraft control failures. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (143-154). New York: Plenum Press.

Fadden, S., Ververs, P., & Wickens, C. D. (1998). Costs and benefits of head-up display use: A meta-analytic approach. *Proceedings of the 42nd Annual Meeting of the Human Factors & Ergonomics Society* (16-20). Santa Monica, CA: Human Factors and Ergonomics Society.

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.

Hoffman, R. R., Crandall, B., & Shadbolt, N. (1998). Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors, 40*, 254-276.

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors, 48*, 196-205.

Hutchins, S. D., Wickens, C. D., Carolan, T. F., & Cumming, J. M. (2013). The influence of cognitive load on transfer with error prevention training methods a meta-analysis. *Human Factors, 55,* 854-874.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science, 5,* 113-153.

Kaber, D. B., Onal, E. & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing 10*, 409-430.

Kessel, C. J. & Wickens, C. D. (1982). The transfer of failure-detection skills between monitoring and controlling dynamic systems. *Human Factors, 24*, 49-60

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an “aid” can (and should) go unused. *Human Factors, 35*, 221–242.

Layton, C., Smith, P. J. & McCoy, C. E. (1994). Design of a cooperative problem-solving system for en-route flight planning: an empirical evaluation. *Human Factors, 36*, 94-119.

Li, H., Wickens, C. D., Sarter, N. B., & Sebok, A. (in preparation). *An investigation of automation degree in assisting robotic arm control.* Technical Report to be submitted.

Lorenz, B., DiNocera, F., Röttger, S., & Parasuraman, R. (2002a). Automated fault-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine, 73,* 886-897.

Lorenz, B., DiNocera, F., Röttger, S., & Parasuraman, R. (2002b). Varying types and levels of automation in the support of dynamic fault management: An analysis of performance costs and benefits. In D. de Waard, K. A. Brookhuis, J. Moraal, and A. Toffetti*, Human Factors in Transportation, Communication, Health, and the Workplace* (517 - 524). Maastricht, the Netherlands: Shaker.

Manzey, D., Reichenbach, J., & Onnasch. L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making, 6,* 57 - 87.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, *47*, 35–49.

Miller, C. A. & Parasuraman, R. (2007). Designing for flexible interaction of humans and automation: delegation interfaces for supervisory control. *Human Factors, 49*, 57-75,

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (201-220). Mahwah, NJ: Lawrence Erlbaum Associates.

Osgood, J. (1949). The similarity paradox in human learning: a resolution. *Psychological Review, 47,* 419–27.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: A review and attentional synthesis. *Human Factors, 52*, 381-410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced “complacency”. *International Journal of Aviation Psychology, 3*, 1-23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misue, disuse and abuse. *Human Factors, 39,* 230-253.

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

Parasuraman, R., Sheridan, T., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering & Decision Making, 6,* 140-160.

Rasmussen, J., & Rouse, W. B. (1981). *Human detection and diagnosis of system failures*. New York: Plenum Press.

Reichenbach, J., Onnasch, L., & Manzey, D. (2011). Human performance consequences of automated decision aids in states of sleep loss. *Human Factors, 53,* 717-728.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage

Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, *52*, 512–523.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors, 49,* 76-87.

Sarter, N. B. (2008). Investigating mode errors on automated flight decks: Illustrating the problem-driven, cumulative, and interdisciplinary nature of human factors research. *Human Factors Golden Anniversary Special Issue, 50,* 506-510.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors, 43*, 573-583.

Seppelt, B. D., & Lee, J. D. (2007) Making adaptive cruise control (ACC) limits visible. *International Journal of Human-Computer Studies, 65,* 192–205.

Sethumadhavan, A. (2009). Effects of automation types on air traffic controller situation awareness and performance. *In Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting* (1-5). Santa Monica, CA: Human Factors and Ergonomics Society.

Sheridan, T. B. (2002). *Humans and automation: Systems design and research issues.* New York: Wiley.

Sheridan, T. B., & Parasuraman, R. (2000). Human vs. automation in responding to failures: An expected-value analysis. *Human Factors, 42,* 403-407.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators.* (Technical Report). Cambridge, MA: Man Machine Systems Laboratory, MIT.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenom. *Journal of Experimental Psychology: Human Learning & Memory, 4,* 592-604.

Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational awareness in aerospace* *operations* (AGARD-CP-478; pp. 3/1-3/17). Neuilly Sur Seine, France: NATO-AGARD.

Wickens, C. D. (2008a). Situation awareness: review of Mica Endsley's 1995 articles on situation awareness theory and measurement. *Human Factors, 50*, 397–403

Wickens, C. D. (2008b). Multiple resources and mental workload. *Human Factors, 50,* 449-455.

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

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R., (2012). *Engineering psychology and human performance* (4th ed.). Pearson.

Wickens, C. D., Hooey, B. L., Gore, B. F., Sebok, A., & Koenicke, C. S. (2009). Identifying black swans in NextGen: predicting human performance in off-nominal conditions. *Human Factors, 51*, 638-651.

Wickens, C. D., Hutchinson, S., Carolan, T., & Cumming, J. (2013). Effectiveness of part task training and increasing difficulty training strategies: A meta-analysis approach. *Human Factors, 55,* 461 - 470.

Wickens, C. D. & Kessel, C. J. (1979). The effects of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions on Systems, Man and Cybernetics, 9*, 24-34.

Wickens, C. D. & Kessel, C. J. (1980). Processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 564-577.

Wickens, C. D., & Kessel, C. J. (1981). Failure detection in dynamic systems. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures*. (155-169). New York: Plenum Press.

Wickens, C. D., Mavor, A. S., Parasuraman, R., & McGee, P. (1998). Airspace system integration: The concept of free flight. In C. D. Wickens, A. S. Mavor, & J. P. McGee (Eds.), *The future of air traffic control: Human operators and automation* (225–245). Washington, DC: National Academy.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics, 23,* 995-1011.

Wright, M. C., & Kaber, D. B. (2005). Effects of automation of information-processing functions on teamwork. *Human Factors*, *47*, 50–66.

Yeh, M., Merlo, J. L., Wickens, C. D., & Brandenburg, D. L. (2003). Head up vs. head down: Costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors, 45*, 390-407.

3. Study II

Operators’ Adaptation to Imperfect Automation – Impact of Miss-Prone Alarm Systems on Attention Allocation and Performance

Linda Onnasch, Stefan Ruff, Dietrich Manzey

Department of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany

Reference

Onnasch, L., Ruff, S. & Manzey, D. (2014). Operators’ Adaptation to Imperfect Automation – Impact of Miss Prone Alarm Systems on Attention Allocation and Performance. *International Journal of Human-Computer Studies*, *72*, 772-782.

3.1 Abstract

Operators in complex environments are often 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 to compensate for unreliability and to maintain overall performance. This study explores how humans adapt to differing alarm reliabilities. Within a multi-task simulation consisting of a monitoring task and two other concurrent tasks, participants were assigned to one of five groups. In the manual control group none of the tasks was supported by an alarm system, whereas the four experimental groups were supported in the monitoring task by a miss-prone alarm system differing in reliability, i.e. 68.75%, 75%, 87.5%, 93.75%. Compared to the manual control group, all experimental groups benefited from the support by alarms, with best performance for the highest reliability condition. However, for the lowest reliability group the benefit was associated with an increased attentional effort, a more demanding attention allocation strategy, and a declined relative performance in a concurrent task. Results are discussed in the context of recent automation research.

Keywords: alarm systems, reliability, miss-prone automation, attention allocation, adaptive behaviour

3.2 Introduction

3.2.1 Alarm Systems

Alarm systems represent a very basic form of automation, typically implemented to gather and analyse information on a certain task in order to inform a human operator about critical states or events, and to support the operator's attention allocation and decision-making. According to Parasuraman et al. (2000), this kind of automation represents the first two stages of their framework model, i.e. automation of information acquisition and information analysis. Information acquisition is automated when an alarm system monitors a single parameter and alerts the operator when critical thresholds are exceeded. If the alarm system is more complex, i.e. if it integrates different variables to detect a possible hazard, it involves both, automation of information acquisition and analysis (Pritchett, 2001). The common characteristic of these two types of information automation is that only cognitive functions related to the sensory perception and evaluation of environmental information are delegated to the automation whereas processes of decision-making and response selection as well as response execution are still left to the human (stages 3 and 4, Parasuraman et al., 2000).

Binary alarm systems are a stereotypical realisation of this widespread technology. The objective of these alarm systems is to support complex supervisory control tasks of operators. Typically, they are implemented in domains like aviation or the process industry where the monitoring of underlying system states and process information constitutes just one of several tasks that have to be performed by operators at the same time. The support provided by alarm systems is mainly enabled by the attention-grabbing properties of alarms which relieve operators from continuous monitoring of a given process while still staying in the loop as alerts inform them when to shift attention to a critical system state (Pritchett, 2001).

Benefits of this type of automation can be described in terms of more efficient task management and prioritisation, as well as reduced operator workload. This in turn leads to a better performance in the task and improved performance in concurrent tasks as operators gain more spare capacities, which can be re-allocated (e.g. Bustamante et al., 2004; Meyer and Bitan, 2002).

However, the proposed benefit of this kind of automation can be off-set when alarm systems do not function properly. The reason for such alarm failures can be found in imperfect sensors and algorithms as well as in a noisy and uncertain world that cannot be interpreted distinctively by the alarm system. Generally, the performance of alarm systems can be described in the framework of signal detection theory (Green and Swets, 1966; Swets, 1964). Following this framework, there are two different errors that can occur and have to be differentiated dependent on the response criterion of the system. First, an alarm system can be miss-prone, i.e. the alarm system can fail to alert the operator by missing critical events. Second, an alarm system can be false-alarm prone. This is the case if it alerts an operator too often as not every alert corresponds to a critical event (Green and Swets, 1966; Swets, 1964). Given these possible failures, operators’ responses to alarms always imply a decision under uncertainty. This decision reflects their assessment of how much they can rely on the alarm function.

3.2.2 Reliance vs. Compliance

According to Meyer (2001, 2004), the explicit distinction between the two kinds of unreliability in human-alarm interaction is important because of their exclusive behavioural consequences on the human part. False alarms may lead to delayed responses towards an alarm as operators know from experience that many of the alarms provided by the system do not correspond to actual malfunction (Getty et al., 1995). In extreme cases, i.e. in cases of high frequencies of false alarms, operators even refuse to respond to an alarm at all (Breznitz, 1984). Misses on the other hand affect operators' monitoring strategies in non-alarm periods. The more critical events are missed by the alarm system, the more operators must shift attention to the alarm-supported task and the raw data to compensate for this unreliability.

Meyer (2001, 2004) therefore characterises operators' behaviour as dependent on the alarm systems' state, i.e. if an alarm is present or not. In this context, compliance refers to operators' response to an alert that indicates a malfunction of the system and is mainly affected by the number of false-alarms emitted by a system. In contrast, reliance describes operators' tendency to rely on the alarm system when it indicates that the monitored process runs properly and the operators accordingly do not have to take evasive action. This latter behavioural tendency represents the major focus of the present paper and shall be addressed in some more detail in the following.

3.2.3 Operators' Adaptation to Imperfect Alarm Systems

According to Lee and See (2004), one of the most important perceivable characteristics for the calibration of reliance on automation (like alarm systems) is the system's reliability. With respect to miss-prone alarm systems, reliability can be described as the percentage of critical events that are correctly indicated by the alarm system. The higher the alarm system's reliability in this respect, the more operators can rely on the alarm and the less they are required to monitor the underlying data by themselves. In contrast, when reliability is low and the occurrence of misses cannot be excluded, operators have to monitor relevant process data more frequently in order to compensate for the alarm system's imperfection and to keep overall monitoring performance high.

Calibration of reliance and compliance therefore can be considered as the result of an adaptive process which develops over time in interaction with an automated system, dependent on the user's experience with the automation's reliability (Lee and See, 2004; Parasuraman and Manzey, 2010).

How and to what extent operators adjust their own monitoring behaviour in case of the availability of (imperfect) alarm systems or other decision support has been addressed in several studies (e.g. Parasuraman et al., 1993; Wickens and Dixon, 2007). However, the results are mixed and provide a somewhat inconsistent pattern of effects. For example, Bailey and Scerbo (2007) examined operators' adaptation to a highly reliable support system. In three sessions, each lasting approximately 100 minutes, participants had to work on a manually controlled flight task while monitoring several simulated aircraft displays for failures. The monitoring tasks were supported by an alarm system that automatically indicated and resolved critical system states. Results indicated that participants' monitoring of the supported task decreased as a function of increasing system reliability, which was set to 87%, 98% and 99.7%, respectively. Participants who were supported by a highly reliable but still not perfect alarm system did detect fewer automation misses and showed increased response latencies to critical events when not alerted by the system, compared to participants who worked with an alarm system with lower reliability. Time-on-task had no effect on these results, i.e. even participants with more system experience and supported by a highly reliable alarm system could not appropriately adapt to automation's imperfection. These findings supported earlier results by Molloy and Parasuraman (1996) who also reported degraded monitoring performance in terms of less miss detection when participants interacted with a highly reliable alarm system. However, they are in contrast to a number of other studies which suggest that operators indeed are very well capable of adapting their own monitoring behaviour to changing reliability levels, suggesting nearly optimum calibration of their reliance on automation reliability (e.g. Parasuraman et al., 1993; Sharma, 1999; Singh et al., 2005; Singh et al., 1997; Wiegmann et al., 2001).

In most of these studies however, the evaluation of monitoring performance was solely based on operator's performance (Bailey and Scerbo, 2007; Parasuraman et al., 1993; Wiegmann et al., 2001). 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. As a consequence it is considered important to always respect the joint human-automation performance while evaluating overall performance benefits or costs associated with this sort of automated support.

In accordance with this approach, Wickens and Dixon (2007) conducted a meta-analysis consisting of 22 studies with varying reliabilities. In contrast to most interpretations of the aforementioned research, they found a positive linear relation between automation's reliability and the joint human-automation performance. That is, even though operators may have tended to miss more critical events when working with alarm systems of high reliability compared to systems with lower reliability, the overall number of *jointly* detected critical events was still higher with highly reliable systems than with lower ones. However, below an alarm system's reliability of 70%, accompanied by a 95% confidence interval, which brackets 65% and 75%, this compensation was associated with disproportional effort, and joint performance even got worse than working with no automation at all. Thus, compensation for unreliability seems to be possible to a certain level only.

This finding is supported by several other studies like, for example, a series of studies conducted by Dixon et al. (e. g. Dixon et al., 2004; Dixon et al., 2007; Dixon and Wickens, 2006). In these studies, Dixon et al. (e. g. Dixon et al., 2004; Dixon et al., 2007; Dixon and Wickens, 2006) compared different levels of reliability of an alarm system supporting monitoring performance in a multi-task environment. They also found certain cost effects on concurrent task performance for alarm system reliabilities at least below 70%. When imperfect alarm reliability was realised by an increased number of misses, operators re-allocated their attention to the alarm-supported task to such extent that a high performance level in the alarm-supported task was maintained. However, concurrent task performance even dropped below the performance of a manual control group without automation support. This drop of performance was explained by a sort of overcompensation effect. The low reliability of the alarm system led to such a decrease in reliance on alarms that participants started to shift more attention than necessary to the alarm-supported task in order to compensate for the imperfection of their system.

Finally, the assumption that operator’s adaptation to imperfect alarm systems might not be perfect - particularly for low reliability systems - is also supported by a study conducted by Wickens et al. (2005). In contrast to the aforementioned studies, Wickens et al. (2005) did not just evaluate possible costs of imperfect reliability on the performance level but also used eye-tracking data to directly evaluate the impact of different reliabilities on visual attention allocation. This additional evaluation level, i.e. eye-tracking data for attention allocation, complies with Moray’s and Inagaki’s (2000) assertion to evaluate operators’ performance not only by fault detection but first and foremost by an analysis of their attention allocation strategies. Participants were required to work on a multi-task scenario based on demands of unmanned air vehicle (UAV) control and several UAV-mission-related tasks that had to be performed concurrently. One of these latter tasks was supported by a binary auditory alarm system that was either perfectly reliable, 60% reliable in terms of misses (miss-prone) or 60% reliable in terms of false alarms (false-alarm prone). Additionally, these groups were compared to a baseline condition in which no automation support for any task was available. Most interesting to the current study was the result that working with the miss-prone automation removed visual attention from the concurrent tasks to the alarm-supported task. In the attempt to maintain adequate performance, participants drew even more attention to the alarm-supported task than in the baseline condition without automation support. Yet, even with this strategy, performance in the alarm-supported task dropped below the baseline condition level.

Summarizing the scope of this research it can be assumed that human operators adapt their behaviour to the characteristics of the automation they are working with. However, there is evidence that this adaptation might not always be appropriate. Studies focussing on human monitoring performance alone suggest that particularly highly reliable alarm systems might lead to miscalibrations of behaviour in terms of an inappropriate withdrawal of attention from the alarm-supported tasks, and an elevated risk of missing critical events. Studies focussing on joint human-system performance specifically point to issues related to low reliable systems (i.e. reliability $< .70$) which might reduce reliance levels to an extent that it becomes even more detrimental for concurrent task performance than working without any automation support.

However, there are two common drawbacks of most of the studies conducted thus far. The first one concerns the relatively extreme levels of automation reliability that were usually compared in those studies, and thus failed to describe the characteristics of adaptation across a whole range of reliability levels. Second, most studies that explicitly varied reliability only concentrated on the state manifestation of reliability effects on human performance, hence excluding the adaptation process itself (some exceptions are Parasuraman et al.,1993 or Bailey and Scerbo, 2007). Although, researchers in the early 90s already argued that system experience has substantial impact on how operators interact with and monitor automation (e.g. Lee and Moray, 1992; Muir, 1987, 1994), only few studies have picked up this claim and focused on reliance development since then. What is known to date is that the adaptation to automation's characteristics seems to proceed fast, and that already single automation failures can have a detrimental impact on users' trust and behaviour (e.g. Bahner, Hüper, Manzey, 2008; Lee and See, 2004; Parasuraman and Manzey, 2010; Manzey, Reichenbach, Onnasch, 2012). Beyond that, only little is known about how these effects develop dependent on different reliability levels, to what extent they are reflected in changes of monitoring strategies, and what the performance consequences are in multi-task environments.

Based on these findings, the goal of the current study was to gain further insight into possible adaptation strategies to alarm systems with respect to different levels of alarm reliability. In contrast to numerous other studies that have concentrated on false alarm-prone automation (e.g. Bliss and Dunn, 2000; Bliss et al., 1995; Lees and Lee, 2007; Wickens et al., 2009), the focus of our study was on miss-prone alarm systems. Even though this kind of error seems to occur less often because designers tend to set sensor thresholds at a very low level (engineering fail safe approach; Swets, 1992), the consequences of missing critical events in safety-related domains are usually more severe than consequences of false alarms. For this reason, it was of special interest if and how operators would compensate for this kind of diagnostic failure.

The task used for the experiment was a multi-task simulation, including three different subtasks. One of these tasks involved a system monitoring task where participants had to monitor different engine gauges for possible failures with or without support of a binary visual alarm system of different reliability. To evaluate participants' monitoring effectiveness, we considered the joint human-automation performance as well as participants' performance in concurrent tasks. In addition, eye-tracking analyses were performed in order to directly assess the impact of alarm system's reliability on participants' attention allocation. By separate analyses of eye-tracking data for periods where alarms were emitted vs. non-alarm periods it was further possible to distinguish between effects of alarm reliability on the level of participant's reliance and compliance.

For the impact of alarm reliability on **performance** we hypothesised:

(1) There is an automation benefit in the alarm-supported task in terms of a superior joint performance of human and alarm system compared to no automation support at all.

(2) Automation benefits in terms of a superior joint performance of human and alarm system compared to no automation support at all are positively related to the alarm system's reliability (Wickens and Dixon, 2007).

(3) Concurrent task performance benefits from highly reliable automation support compared to the manual control condition. However, these benefits decrease with decreasing alarm reliability over time because participants start to reallocate attention to the alarm-supported task to compensate for automation's imperfection.

In extreme cases, i.e. interacting with an automation with a reliability below the critical cut-off of 70%, this adaptation of attentional reallocation should even lead to cost effects in terms of a degraded performance compared to working with no automation support at all (Dixon et al., 2007; Dixon and Wickens; 2006; Rovira et al., 2007; Wickens and Dixon, 2007; Wickens et al., 2005).

For participants' **visual attention allocation**, operationalised by eye-tracking measures, we expected:

(4) Participants supported by an alarm system of sufficient reliability invest less attentional resources in system monitoring compared to working with no automation support.

(5) Participants adapt their own monitoring of engine gauges to the alarm systems' reliability over time.

Participants working together with relatively reliable automation support should decrease their own monitoring with growing system experience whereas participants supported by an unreliable automation should increase monitoring of the underlying data (engine gauges).

(6) In interaction with alarm reliability below 70% participants' attention allocation is not distinguishable from attention allocation when working manually on this task as compensation for unreliability becomes inefficient (Wickens et al., 2005).

(7) Because we operationalised reliability only by misses of the alarm system, differences in participants' attention allocation primarily emerge during non-alarm periods, reflecting effects on participants' reliance.

No or only little differences were expected for visual attention effects in direct response to alarms, which would reflect the level of compliance and which was expected to be high for systems that did not commit false alarms.

3.3 Method

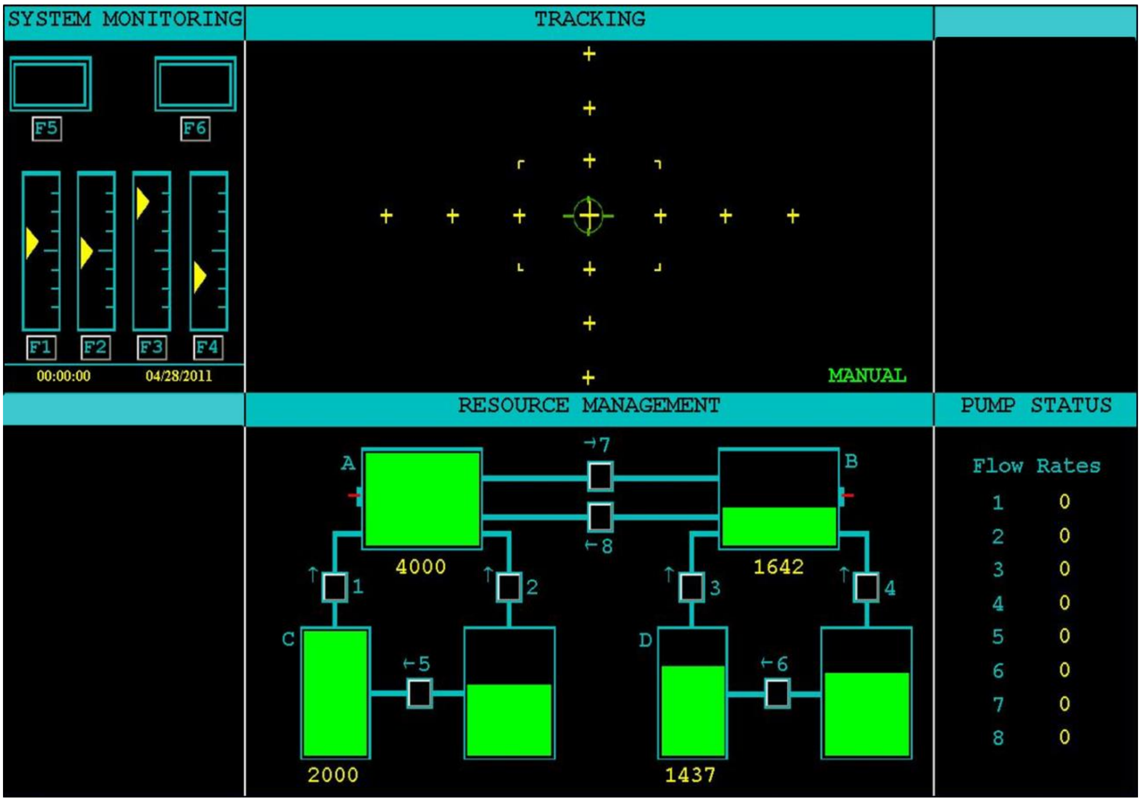
3.3.1 Participants

The number of participants was defined based on a power analysis (GPower 3.1, for details see e.g. Buchner et al., 1997). A total of 65 students from the faculty of mechanical engineering and transport systems (18 female, 47 male) ranging in age from 19 to 32 ($M = 23.6$, $SD = 2.3$) participated in partial fulfilment of course requirements. None of the participants had prior experience with the flight simulation task used in the study. Participation was voluntary (other alternatives for fulfilment of course requirements were available) and could be cancelled anytime.

3.3.2 Task and Apparatus

As experimental task the most recent version of the *Multi-Attribute Task Battery* (MATB; Miller, 2010) was used. It was directly based on the original version developed by Comstock and Arnegard (1992) which was used in previous research (e.g. Parasuraman et al., 1993). All main functionalities including the interface corresponded to the original version. Only the programming environment has been changed (MatLab instead of QBasic) which made it easier to implement experimental modifications.

The MATB is a multi-task flight simulation consisting of a two dimensional compensatory tracking, engine-system monitoring, fuel resource management, communications, and scheduling. In the present study, only the compensatory tracking, the resource management, and the system monitoring were implemented and had to be performed concurrently. The user interface of the MATB used in the present study is shown in Figure 1.



Flow	Rates
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0

Figure 1. MATB as used in the current study with the compensatory tracking in the upper middle position, the resource management beneath and the system monitoring in the upper left display corner.

In the two-dimensional compensatory tracking task participants are required to keep a randomly moving cursor in the centre target position by applying appropriate control inputs via joystick. In the resource management task participants must compensate for fuel depletion by pumping fuel from four supply tanks into two main tanks.

The system monitoring task was most important for the current research. It consists of four vertical engine gauges with moving pointers that participants must monitor for abnormal values that occur randomly. As long as all engines function properly, the pointers fluctuate by chance within a fixed range around the centre value of the gauges. However, in case of a malfunction the pointer of the gauge for the affected engine suddenly shifts upwards or downwards by two gauge units and starts to fluctuate around this new position. These deviations must be detected by participants and reset by a corresponding key press. If a malfunction is not detected within 10 seconds the gauge resets automatically and the event is logged as an event missed by the participant.

Dependent on the task configuration, this system monitoring task has to be performed manually or with support of a binary master alarm system. In the latter case, a visual red alert appears above the gauges whenever the alarm system detects a parameter deviating from its nominal value. Nevertheless, the identification of the affected gauge and the corresponding reset of the parameter still have to be performed manually by participant. According to the stages and levels framework of automation proposed by Parasuraman et al. (2000), this type of alarm system can be classified as a stage 1 automation (information acquisition).

The MATB was presented in front of the participant on a 20 inch monitor that was equipped with a remote eye-tracking system (RED system, SensoryMotoric Instruments, Germany). This latter system enabled to sample gaze movements during task performance with a sampling rate of 120 Hz. Based on these data, gaze fixations in different areas of interest (AOI, see definition below) were automatically recorded.

3.3.3 Design

The study used a two factorial design. The first factor (Group) was defined as a between-subject factor and consisted of four experimental groups and one manual control group. The four experimental groups differed with respect to the reliability of the alarm system participants worked with in the monitoring task. The alarm reliabilities were set to 68.75%, 75%, 87.5%, and 93.75% by varying the number of critical signals that were missed by the alarm system. The two lowest reliability levels (68.75% and 75%) were chosen in reference to the result of the meta-analysis of Wickens and Dixon (2007) which suggests that a reliability level around .70 represents an important cut-off value which needs to be exceeded before automation support might become beneficial for joint human-system performance compared to conditions without automation support. The two highest reliability levels were realised to compare the results to findings from previous studies and to include reliability levels quite close to realistic scenarios (Bagheri and Jamieson, 2004; Parasuraman et al., 1993). In the manual control group there was no automation support at all, i.e. participants had to detect all malfunctions reflected by parameter deviations in one of the four gauges without the support of an alarm system.

The second factor (Block) was defined as a within-subject factor and was included to gain further insight on how participants' adapt their attention and performance over time in response to the alarm system's reliability they were working with. Every participant had to perform the three concurrent tasks of the MATB for three 10-minute blocks. A total of 16 critical events occurred in the monitoring task during each block which had to be detected by the alarm system or the participant, respectively. The resulting 5 (Group) x 3 (Block) design is shown in Figure 2.

		Block		
		block #1	block #2	block #3
Group	manual		n = 13	
	68.75%		n = 13	
	75.00%		n = 13	
	87.50%		n = 13	
	93.75%		n = 13	

Figure 2. 5 (Group) x 3 (Block) study design

A somewhat more complex design was used for supplementary analyses of effects of reliability on visual attention allocation in phases where alarms were present vs. phases where alarms were not present. The beginning of alarm phases could be identified by the visual red alert that appeared to inform participants about an abnormal system state. The end of these phases was defined by participants' appropriate reaction to the alarm or, if participants did not react, the maximum time the failure was present, i.e. 10 seconds. These supplementary analyses involved the four alarm-supported groups as between-subjects factor, the block factor (within-subject) and a third factor representing alarm vs. non-alarm periods (within-subject). The resulting 4 (Group) x 3 (Block) x 2 (Alarm State) design allowed a test of the hypothesis that differences in reliability of the alarm system would affect attention reallocation during non-alarm periods only, reflecting effects on reliance on the automation but not compliance (Hypothesis 7).

3.3.4 Dependent Measures

To investigate the impact of the experimental factors on the perceived alarm reliability (manipulation check) as well as on performance and visual attention allocation, three different categories of dependent measures were sampled and analysed.

A visual-analogue scale assessed the perceived reliability. Participants provided ratings to the question “How reliable was the system you worked with on a scale ranging from 0% to 100%.

Performance measures were defined for all three tasks of the MATB participants had to perform concurrently and collected for each 10 minute block separately. For the system monitoring task, *percentage of detected system failures* was defined as the percentage of all engine failures detected correctly by the human operator (control condition) or the human and alarm system together (joint performance in the alarm conditions).

For the tracking task as well as the resource management task the *root mean squared errors* (RMSE; Parasuraman et al., 1993; Prinzel et al., 2001; Singh et al., 1997) were calculated. The RMSE for the tracking task was calculated as a measure of mean deviation from the central target position, based on deviation data sampled at an interval of 5 seconds. The RMSE for the resource management task was calculated in relation to an optimal tank level, which had to be maintained in both main tanks. Fuel levels were sampled and RMS errors computed for each 5-second period.

Visual attention allocation was measured by means of eye-tracking. Specifically, the *relative fixation time* for different pre-defined areas of interest (AOI) was assessed. For this purpose, three different AOIs (specified by pixel areas) were defined before the experiment started. These AOIs corresponded to the three different tasks participants had to perform: compensatory tracking, resource management, and system monitoring (see Figure 1). Fixations were defined by a minimum duration of 80 ms and a maximum dispersion in this time of 100 pixel. Relative fixation time was defined as the time participants fixated an AOI relative to the overall fixation time, i.e. sum of times any one of the AOIs was fixated.

3.3.5 Procedure

Following a demographic questionnaire, an instruction on the MATB, and an initial calibration of the eye-tracking system, participants were familiarised with performing the three different tasks in a 10 minute practice block. They were instructed that all three tasks would be of equal importance, and that they should work on all tasks concurrently with equal priority. Afterwards, they were randomly assigned to one of the five groups. Participants in the four experimental groups were introduced to the function of the alarm system. Specifically, they were told that the alarm system would not be perfectly reliable and that therefore, they may not fully rely on it. However, no concrete reliability information was provided. Then, the experiment started consisting of three 10 minute blocks. Prior to each block the eye tracker was re-calibrated. The perceived reliability of the alarm system was assessed in the experimental groups after the second block. The experiment ended with the debriefing of participants.

3.4 Results

In the following, the results are presented separately for subjective measures, performance, and eye-tracking data. The description of results focusses on effects of reliability (factor Group) and/ or possible interactions with time-on-task (factor Block), indicating adaptive processes.

3.4.1 Perceived Reliability

A univariate between-subjects ANOVA contrasting the four experimental conditions with automation support of different reliability revealed that mean ratings of perceived reliability differed between these experimental groups in a meaningful manner ($M_{68.75\%} = 66.77\%$, $M_{75\%} = 72.38\%$, $M_{87.5\%} = 80.08\%$, $M_{93.75\%} = 87.08\%$), F(3, 51) = 6.11, p < .002.

Further t-tests were performed in order to analyse whether perceived ratings differed from the actual reliability. Because no differences were expected, α was adapted to a 20% level for these analyses (null-hypothesis testing). Results showed that participants in the two highest reliability conditions systematically underestimated the actual reliability (87.5%: t(12) = -3.29, p < .007; 93.75%: t(12) = -3.09, p < .01). No differences between actual and perceived reliability were found for the 68.75% and 75%

reliability condition (68.75%: $t(12) = -.48$, $p = .63$; 75%: $t(12) = -.52$, $p = .61$). This finding is in line with previous research (Wiczorek and Manzey, 2010; Wiegmann et al., 2001; Wiegmann and Cristina, 2000) indicating a systematic bias of under- and overestimation, respectively, for extreme levels of reliability. Nevertheless, the overall pattern of results confirms that our manipulation had worked successfully as the perceived reliabilities were systematically related to the actual ones and significantly differed between the experimental conditions.

3.4.2 Performance Measures

3.4.2.1 Monitoring Task

Performance measures were analysed in two steps according to the different hypotheses. The first step addressed the testing of our hypothesis which postulated an alarm-support benefit in the monitoring task compared to no alarm-support at all (Hypothesis 1).

For this purpose, the *percentage of detected system failures* was analysed with a 5 (Group) x 3 (Block) ANOVA. The corresponding data, i.e. detection rates for all experimental groups and the manual control group across blocks, are shown in Figure 3. As expected, there was a clear alarm-support advantage reflected in a higher percentage of detected system failures by human and automation together in all alarm-supported groups, compared to the manual control group ($F(4, 60) = 10.36$, $p < .001$, $\eta^2 = .40$). Averaged across blocks, participants of the control group only detected 73.23% of all failures. In contrast, participants in the experimental group with the least reliable alarm system already detected 90.70% of all failures, and this number increased systematically with increasing reliability of alarms ($M_{75\%} = 92.46$, $M_{87.5\%} = 93.26$, $M_{93.75\%} = 95.83$). This difference between automation-supported groups and the manual control group was statistically supported by post hoc analyses using Scheffe' tests. Analyses revealed that the manual control group detected significantly less system malfunctions compared to any of the alarm-supported groups ($p_{manual-68.75\%} < .003$; $p_{manual-75\%} < .001$; $p_{manual-87.5\%} < .001$; $p_{manual-93.75\%} < .001$). No differences occurred between the alarm-supported groups (all $p > .05$). Additionally, an interaction of reliability with participants' time-on-task was found, Group x Block interaction effect, $F(8, 120) = 2.37$, $p < .03$, $\eta^2 = .13$. Whereas all conditions showed an improved performance across blocks, the extent of this performance increase was different for the five groups. The largest increase in detected system failures over time was observed for the manual control group. In this condition, no alarm system support was available. Still, participants had to adapt to the underlying system characteristics and get familiar with the error rate in the monitoring task to perform adequately. As becomes evident from Figure 3, this form of adaptation was comparable to a similar, albeit weaker trend of participants' behaviour in the group working with the least reliable alarm system. Compared to the other conditions with alarm support, this group showed the worst performance at the beginning, but participants adapted their behaviour to the characteristics of the alarm system over time and were able to compensate effectively for its unreliability. However, this latter difference between the alarm-supported groups did not become significant in an additional 4(Group) x 3(Block) ANOVA, comparing the alarm-supported groups only. For this analysis neither the expected effect of Group (F = 1.69), nor a Group x Block interaction effect (F = 1.16) emerged (contradicting Hypothesis 2).

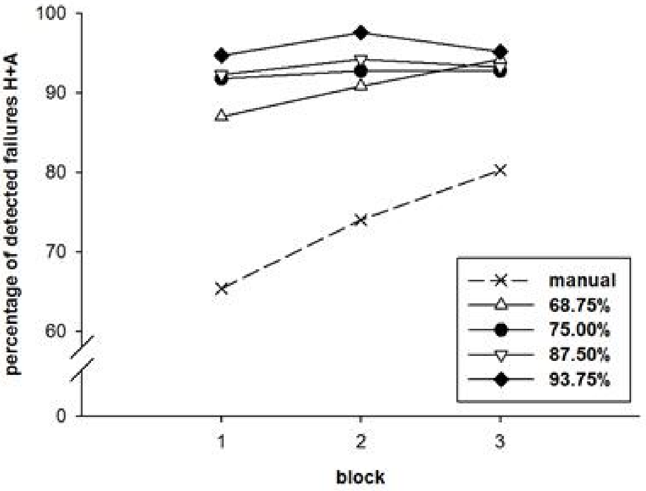


Figure 3. Effect of alarm reliability on detected system failures - human + alarm system.

3.4.2.2 Concurrent Tasks

Following the same statistical approach as for the monitoring task, performance in the concurrent tasks was analysed in two steps. We expected that compared to higher reliability levels, working with the least reliable alarm system would negatively affect concurrent task performance because participants would rely to a lesser extent on the proper functioning of the alarm support (Hypothesis 3). More specifically, it was expected that concurrent task performance of the 68.75% reliability group would not be better than performance in the manual control group, i.e. a condition with no automation support at all.

For concurrent *tracking task* performance the 5 (Group) x 3 (Block) ANOVA revealed a significant Group x Block interaction, $F(8, 120) = 3.59$, $p < .002$, $\eta^2 = .19$. Essentially the same pattern of effects was also observed when comparing the alarm-supported groups only by a 4 (Group) x 3 (Block) ANOVA, with a significant interaction effect of Group x Block, $F(6, 96) = 4.98$, $p < .001$, $\eta^2 = .23$.

As can be seen in Figure 4, contrary to our expectations, participants in the 68.75% reliability group started at a very high performance level reflected in a smaller mean tracking error than in all other groups ($M_{manual} = 131.78$, $M_{68.75\%} = 117.58$, $M_{75\%} = 136.05$, $M_{87.5\%} = 144.57$, $M_{93.75\%} = 137.76$). However, whereas participants of the other groups showed a considerable performance improvement over time, mean performance of participants in the 68.75% reliability condition declined across the three blocks. This eventually led to comparable performance levels for all groups in block #3 ($M_{manual} = 124.62$, $M_{68.75\%} = 126.94$, $M_{75\%} = 126.78$, $M_{87.5\%} = 127.58$, $M_{93.75\%} = 129.55$). This finding provides some indirect support for our hypothesis. In contrast to all other alarm-supported groups, participants working with the lowest reliable alarm system were only able to protect their performance in the monitoring task across time at the expense of compensatory decrements in concurrent task performance.

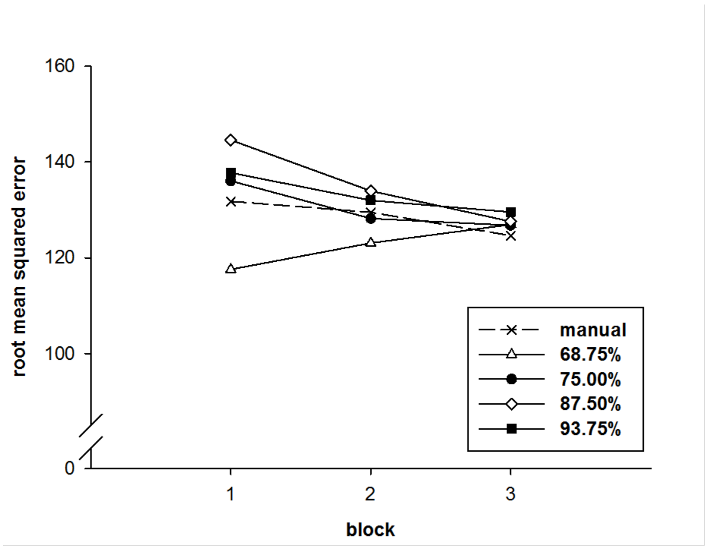


Figure 4. Effect of alarm reliability on performance in the concurrent tracking task (higher values represent greater deviations).

For the *resource management task* neither a main effect of Group nor a Group x Block interaction emerged (all F<1.0). Only a Block effect became significant independent of whether all groups were considered in a 5(Group) x 3(Block) ANOVA, $F(1.2, 75.69) = 5.02$, $p < .03$, $\eta^2 = .07$, or the analysis was only conducted for the four experimental groups with alarm support, $F(1.5, 73.47) = 4.31$, $p < .03$, $\eta^2 = .08$. With increasing time-on-task all groups achieved better results reflected in a decreased mean RMSE.

3.4.3 Visual Attention Allocation

3.4.3.1 Overall Monitoring Effects for the different AOIs

Figure 5 illustrates the results for the mean *relative fixation times* on the three different AOIs, i.e. monitoring task (left panel), tracking task (middle panel) and resource management task (right panel).

For the *monitoring task*, participants in the two highest groups (93.75% & 87.5%) showed relatively short but stable mean *fixation times* across blocks. This effect was expected because these participants could rely to a high degree on the alarm system. Stable mean fixation times across blocks also were found for the 75% reliability group, albeit on a somewhat higher level. In clear contrast to these three groups, a considerable increase of mean fixation time through blocks was found for both, the manual control group as well as the group working with the lowest reliable alarm system (Figure 5, left panel). Analysed by a 5 (Group) x 3 (Block) ANOVA these findings were statistically supported by a significant Group x Block interaction, $F(7.14, 107.13) = 2.46$, $p < .03$, $\eta^2 = .14$.

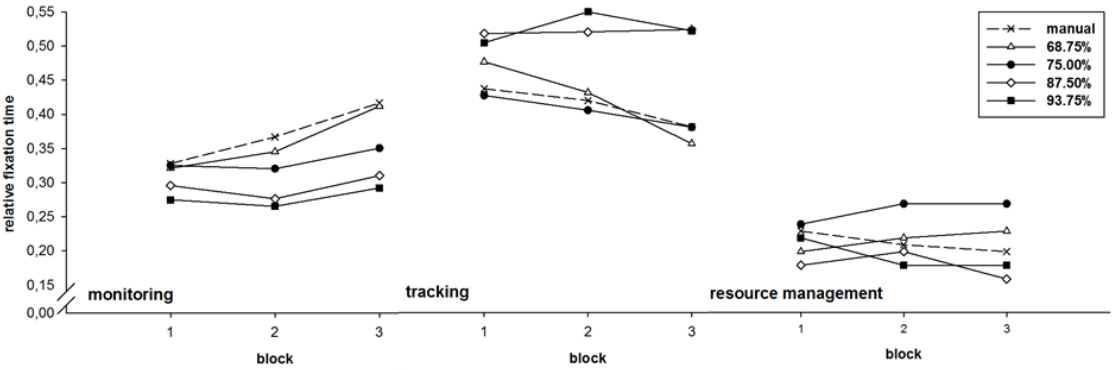


Figure 5. Effect of alarm reliability on the relative fixation time; AOI from left to right: monitoring, tracking and resource management.

Results for the monitoring task were mirrored in the *relative fixation times* for the *tracking task* (Figure 5, middle panel). Directly inverse to the findings for the monitoring task, the 93.75% and the 87.5% reliability groups had the longest fixation times on tracking which only marginally changed over time. For the other groups, a considerable decrease of fixation times across blocks was found which was most substantial for the 68.75% reliability condition and indicated a successive re-allocation of visual attention away from the tracking task over time. The 5(Group) x 3 (Block) ANOVA revealed a significant main effect of Group ($F(4, 60) = 2.64$, $p < .05$, $\eta^2 = .15$), moderated by a Group x Block interaction effect, $F(6.75, 101.29) = 3.62$, $p < .003$, $\eta^2 = .19$.

Finally, mean relative times of fixation for the *resource management task* did not show a clear pattern of effects. The 5 (Group) x 3 (Block) ANOVA did not reveal a main effect of Group ($F= 1.55$), however, the Group x Block interaction became significant, $F(7.16, 107.39) = 2.14$, $p < .05$, $\eta^2 = .12$. As becomes evident from Figure 5 (right panel), relative fixation times showed a slight increase across blocks for the two conditions with the lowest reliable alarm systems, and a reverse trend for the other three groups.

In summary, the pattern of effects for relative fixation times on the three different tasks is in accordance with our hypothesis that alarm reliabilities affected the allocation of visual attention. Specifically, the results point to a successive re-allocation of attention over time, away from the tracking task to the monitoring task. Re-allocation emerged in a very similar way in both, the control condition without automation support and the condition with support of the lowest reliable alarm system.

3.4.3.2 Specific Effects for Alarm and Non-Alarm Periods

As our alarm systems were miss-prone it was expected that they would primarily affect the reliance of participants in the alarm systems' function but not their compliance. Accordingly, it was expected that possible effects of alarm reliability on visual attention allocation would only emerge during periods when no alarm was present (non-alarm periods). During these non-alarm periods participants should allocate more attention to the monitoring task, the less they relied on the proper functioning of the alarm system. I.e. if participants expected that the alarm system could miss critical system states they should reallocate their attention from the other two concurrent tasks to the alarm-supported monitoring task. In contrast, no differences were expected for visual attention allocation in direct response to alarms which never represented false alarms. For the analysis of this presumed effect only the alarm-supported groups were considered, as a differentiation of these periods was not possible for the manual control group who worked without alarm system.

Figure 6 shows mean relative fixation times for all groups across blocks, separated for the three tasks (from left to right), and periods with and without alarm (upper vs. lower panel).

Results for the *monitoring task* revealed that the pattern of effect found in the overall analysis reported above, i.e. an increase in *relative fixation time* across blocks only in the control group and the group working with the lowest reliable alarm system, was exclusively related to non-alarm periods (Figure 6, upper left panel). In contrast, a decrease of mean fixation times across blocks emerged in all groups during alarm periods (Figure 6, lower left panel). In the analysis of these data by a 4(Group) x 3(Block) x 2(Alarm State) ANOVA this was reflected in a significant main effect of Alarm State, $F(1, 48) = 52.44$, $p < .001$, $\eta^2 = .52$, which was moderated by a significant Alarm State x Block interaction, $F(1.74, 83.81) = 27.86$, $p < .001$, $\eta^2 = .36$. Furthermore, the significant main effect of Alarm State indicated that mean relative fixation times for the monitoring task were higher during alarm vs. non-alarm periods, i.e. higher when an alarm prompted the participants to visually analyse which of the four gauges indicated a failure.

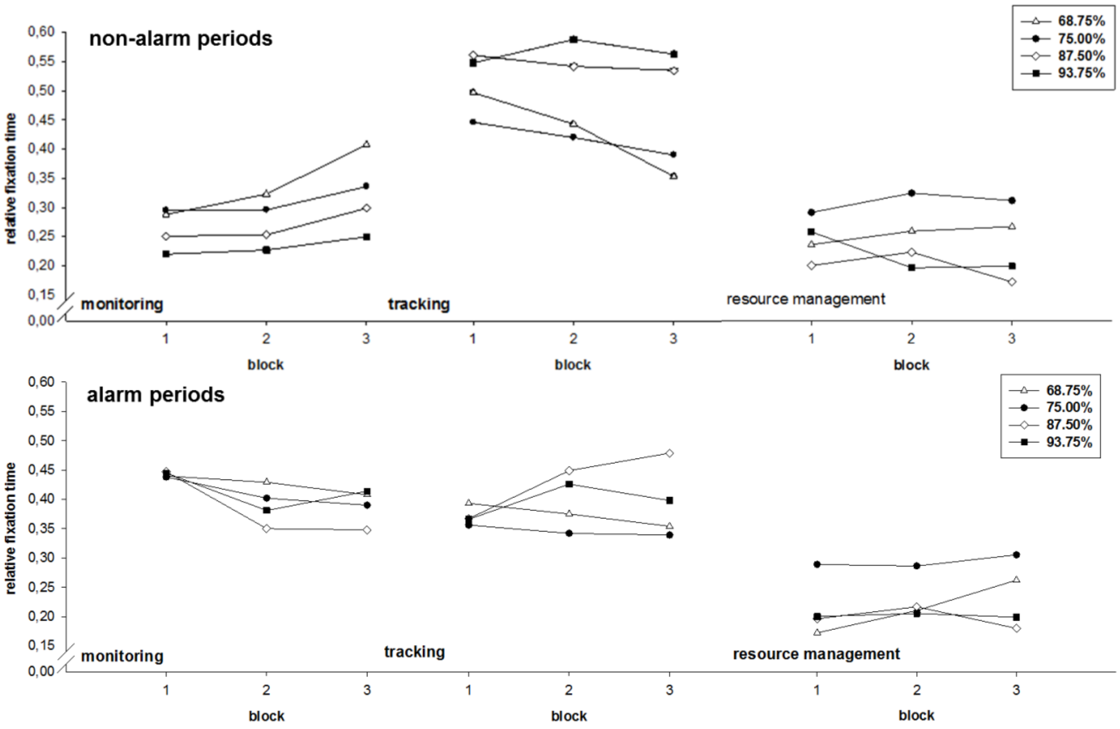


Figure 6. Effect of alarm reliability and alarm state (upper panels non-alarm periods, lower panels alarm periods) on the relative fixation time; AOI from left to right: monitoring, tracking and resource management.

The results for the *tracking task*, separated for alarm and non-alarm periods, are illustrated in the middle panel of Figure 6. Again the effects for non-alarm periods equalled the effects reported above for the overall analysis. In these periods the participants of the two groups with highest alarm reliability had the longest mean fixation times and showed a constant monitoring pattern. In contrast, a continuous decrease of relative fixation times was found for participants in the two lower reliability conditions (see upper panel). This separation of groups became also evident in alarm periods, although in a slightly different pattern (see lower panel). All groups spent comparable time looking at the tracking task in block #1. However, with on-going adaptation to the reliability of the alarm system, the 93.75% and the 87.5% groups spent more time on this task even when a failure in the monitoring task was present. The two groups working with the less reliable alarm systems slightly decreased their monitoring time on tracking in alarm periods. The statistical equivalents to these findings are presented in Table 1.

Table 1. Results of the three-factorial ANOVA for the AOI Tracking Task.

	F-Statistic	**p-value**	**Partial Eta-Squared**
Group x Block	$F(4.76, 76.30) = 4.21$	$p < .003$	$\eta^2 = .20$
Alarm-State	$F(1, 48) = 62.30$	$p < .001$	$\eta^2 = .56$
Group x Alarm-State	$F(3, 48) = 3.66$	$p < .02$	$\eta^2 = .18$
Alarm-State x Block	$F(1.76, 84.78) = 15.10$	$p < .001$	$\eta^2 = .23$
Group x Alarm-State x Block	$F(5.29, 84.78) = 2.58$	$p < .03$	$\eta^2 = .13$

Results for the *resource management* task again revealed that the pattern of effects found in the overall analysis, i.e. an increase in fixation time across blocks for the two lowest reliability conditions, and a reverse effect for the other two conditions, was exclusively related to non-alarm periods, $F(4.70, 75.28) = 2.77$, $p < .03$, $\eta^2 = .14$ (Figure 6, upper right panel). Moreover, when an alarm was present, the resource management was less fixated than in non-alarm periods, $F(1, 48) = 4.39$, $p < .05$, $\eta^2 = .08$; $M_{alarm}= 0.205$, $M_{no\ alarm}= 0.222$. This separation was enforced with ongoing time-on-task, $F(2, 96) = 6.48$, $p < .003$, $\eta^2 = .11$ (Figure 6, right panel).

In summary, results from the monitoring AOI supported the assumption that the attention re-allocation, related to different reliability levels, was only observable in non-alarm periods (hypothesis 7). However, for the tracking task the alarm system's reliability not only affected attention allocation in non-alarm periods but also in alarm periods. Eye-tracking data from the AOI resource management revealed a difference between participants' attention allocation in non-alarm and alarm periods but no interaction of Alarm State and Group. Therefore, results did not fully support Hypothesis 7.

3.5 Discussion

The main objective of this study was to investigate to what extent human operators adapted their visual attention allocation and multi-task performance to different reliability levels of miss-prone alarm systems.

First hypotheses (1-3) were stated with regard to effects of different alarm reliabilities on performance in both, the alarm-supported task as well as other concurrent tasks. Based on previous studies we specifically assumed that participants' performance as well as attentional demands would benefit from an automation support that is fairly reliable, i.e. at least 70% (e.g. Dixon et al., 2007; Rovira et al., 2007; Wickens et al., 2005). Below this reliability, automation support is not expected to be helpful as there is some evidence that a reliability of approximately 70% (accompanied by a 95% confidence interval) represents a critical boundary below which manual compensation strategies would not be effective any more (e.g. Wickens and Dixon, 2007). In this case, we supposed at least performance in concurrent tasks to suffer because participants would start to re-allocate attention to the automation-supported task and monitor the underlying data by themselves to compensate for unreliability. The results of the present study support most of these assumptions.

Considering the results for the performance data first, we found a clear automation benefit in the alarm-supported task in terms of joint human-automation performance compared to working with no automation support at all supporting Hypothesis 1. This was true for all groups that worked with alarm system support. Whereas the manual control group only detected around 70% of engine malfunctions, detection rates increased with alarm-support even in the lowest reliability condition up to 90%. This effect seemed to be an overall automation benefit as differences between the alarm-supported groups did not become significant (contradicting Hypothesis 2). Therefore, the automation benefit was only attributable to the difference between alarm-supported groups on the one hand and the manual control group on the other. This result revealed that all participants in the alarm-supported groups adapted to differing reliability levels in a very effective way. This (non-)finding indicated that participants' adaptation was even more successful than we would have assumed based on findings by Wickens and Dixon (2007) which showed that higher reliability levels still led to significantly improved performance compared to lower reliability levels in the automation-supported task. One reason for these deviant findings might be due to the operational definition of reliability in our study. Whereas Wickens and Dixon (2007) included studies to the meta-analysis that operationalised (un-)reliability by misses and/ or false alarms, we defined reliability by misses only. According to Meyer (2001, 2004) false alarms mainly impact participants' compliance with the automation. As a consequence, false alarms lead to a degraded performance in the automation-supported task as participants start to ignore alarms (Meyer, 2001, 2004). This could explain why Wickens and Dixon (2007) found performance decrements in the automation-supported task when reliability was low. Misses, on the other hand, affect participants' reliance on the alarm system. Because of the frequently missed critical states participants start to monitor the underlying data to compensate for the alarm system's unreliability. This adaptation should not and in fact did not affect performance in the alarm-supported task. However, following Meyer (2001, 2004), concurrent task performance should be negatively affected by this change of attentional focus to the aided task. Therefore, participants' adaptation of monitoring strategies to compensate for the unreliability of the alarm systems was expected to lead to differences in concurrent task performance between the different groups.

This was addressed by our third hypothesis. According to previous studies (e.g. Dixon et al., 2007; Dixon and Wickens; 2006; Rovira et al., 2007; Wickens and Dixon, 2007; Wickens et al., 2005) we expected that, compared to higher reliability levels, lower alarm reliabilities should result in significant performance decrements in the concurrent tasks because participants' reliance on the alarms would decline and induce a re-allocation of attention. Following Wickens and Dixon (2007), we especially assumed that working with an alarm system with a reliability of less than 70% might be even more detrimental to performance than working with no automation support at all. This assumption was at least indirectly supported by our findings.

Although concurrent tracking task performance in the 68.75% group was better than the one of all other groups in the first block, participants in this group were the only ones who could not protect their performance over time but showed a considerable decline across blocks. As we only told participants that the alarm system would not be perfectly reliable, but gave no precise information, the first block was especially important to participants to gain experience with the system and to start to adapt their behaviour to the alarm system's (un-)proper functioning. It seems that participants working with the least reliable alarm system initially spent more time on the concurrent tracking task than on the alarm-supported monitoring task, which resulted in superior results compared to the other conditions. However, with increasing experience they started to realise the limitation of their alarm system and changed their behaviour accordingly by re-allocating their attention away from the tracking task. The other aided groups also started to adapt to the alarm system's characteristics. Because these alarm systems were more reliable, adaptation proceeded the other way, i.e. in favour of the concurrent tracking task, as these groups recognised that they could rely more on their automation support. These diverging adaptation characteristics of the lowest alarm-supported group and conditions with a more reliable alarm-support eventually led to comparable performance levels in the last block.

However, our far-reaching hypothesis that working with the least reliable alarm system would impair concurrent task performance even more than working without automation support was not supported by the data. This might be related to the fact that, contrary to our expectations, the provision of alarm support did not lead to obvious benefits in concurrent task performance in any of the alarm-supported groups. That is, even in the groups with highly reliable alarm systems, the participants were not able to make use of this support in terms of improved concurrent task performance.

One reason for this finding could be the overall high task load involved in performing the MATB. In contrast to, for example, Dixon et al. (2007) or Rovira et al. (2007) who have reported automation benefits for concurrent task performance, participants had to work on three instead of two concurrent tasks. Additionally, the MATB compensatory tracking has high visual attentional demands as it needs continuous control inputs since even short interruptions of control lead to great deviations from the centre target position. Given this, it might not be too surprising that even a reliable alarm support for the monitoring task has not led to better concurrent task performance in our study because participants already performed at their maximum; the tasks were not sensitive to changes in attention allocation. Yet, this is a post-hoc explanation and cannot be fully proved by the present data.

The most direct insights in the nature of adaptation to alarm systems of different reliability are provided by the effect of alarm reliabilities on participants' attention allocation strategies reflected in the eye-tracking data. These data were collected in order to directly capture possible effects of the experimental conditions on allocation of visual attention which might help understanding effects on performance. Indeed the analyses of eye-tracking data suggest that the effects of alarm support first and foremost become evident in their effects on attention allocation (supporting Hypothesis 4).

As expected, participants in the groups with the highest reliable alarms allocated least attention to the monitoring task, followed by the two groups with the less reliable alarm systems and the manual control group (supporting Hypothesis 5).

A comparison of the eye-tracking pattern between the alarm-supported groups and the manual control group further revealed, that the participants of the 68.75% group allocated as much visual attention to the supported task as the manual control group, i.e. behaved as if no automation support were available (supporting Hypothesis 6). It reveals that participants working with the least reliable alarm system were able to compensate for the imperfection of their alarm system on a performance level but only at the expense of a highly demanding attention allocation strategy and a reallocation of attention away from the concurrent tasks which eventually led to the relative performance decline for the tracking. These results are in line with previous findings by Wickens et al. (2005) who also showed that miss-prone automation led to a reallocation of visual attention away from other tasks to the raw data in order to compensate for unreliability. Furthermore, our findings provide some more support for the assumption of a critical reliability cut-off around 70% below which automation support cannot be considered as helpful anymore (Wickens and Dixon, 2007). Albeit we could not entirely confirm a detrimental effect of reliability below 70% on the performance level, the costs for compensation became directly evident when considering the distribution of visual attention. Although the least reliable alarm system still detected 68.75% of all system malfunctions it obviously was not considered to be of much help and did not reduce participants' attentional demands of this task compared to performing it with no automation support at all.

Our last hypothesis (Hypothesis 7) was based on Meyer's assumption (2001, 2004), that misses of an alarm system mainly affect participants' attention allocation in non-alarm periods and have no effect on their visual attention in alarm periods. Regarding the non-alarm periods, this assumption was completely confirmed. The effects found in the overall analysis exactly mirrored participants' attention allocation in non-alarm periods, i.e. the overall effects were mainly due to these periods. This was true for attention allocation on all three concurrent tasks. We could confirm that working with the least reliable aid in terms of misses led to a reallocation of attention away from the tracking and resource management to the alarm-supported monitoring task in the attempt to compensate for unreliability and to maintain performance on this task. In contrast, groups working with more reliable alarm systems maintained the initial level of attention to the supported task and overall focused more on the resource management and tracking.

In alarm periods attention allocation to the supported monitoring task did not seem to be much affected by reliability of the alarm systems. All groups slightly reduced their attention to this task over time but no impact of different reliability levels became evident. This was in line with our assumption, which assumed that only reliance on automation would be affected by a miss-prone alarm system and not compliance (Meyer, 2004).

In conclusion, the current study provides further insight in the adaptation strategies of humans in relation to automation's reliability, one of the most important perceivable characteristics of automation (Lee and See, 2004). The additional value compared to previous studies originates from the level of detail in design and analysis as most of the previous studies only compared two very extreme reliability levels (e.g. Dixon et al., 2007, 2006; Rovira et al., 2007). Furthermore, the analysis of eye-tracking data provided more detailed insight into the impact of alarm systems on attention allocation as compared to the consideration of just performance measures in previous studies.

With regard to practical implications, results are certainly not applicable to high risk work domains like aviation where only alarm systems are used that are optimized in reliability with respect to avoidance of misses and, thus, if ever typically are false-alarm prone. But in other domains like quality control inspection in the manufacturing industry comparable reliability levels even in terms of miss-prone alerting systems, can be found. In this case the finding of a critical reliability cut-off should be taken into account when considering the implementation of such systems. Even though consequences might not be apparent in the beginning, the cognitive effort of operators needed to compensate for the imperfect reliability of such systems could lead to severe problems in the long term, like complete performance breakdowns in the automation-supported task or an overall performance decrease when operators are responsible for multiple concurrent tasks.

3.6 Limitations

Regarding possible limitations of the current study, two aspects should be discussed which might limit the generalisation of results. First, given the fact that participants in the current study only had to work for 30 minutes on the tasks, the results could possibly underestimate some of the observed effects. Especially, the compensation strategies for the alarm's unreliability in order to maintain a high monitoring performance may be difficult to maintain over prolonged periods of time. Ultimately, in terms of cognitive exhaustion, this overexertion might even lead to a complete performance breakdown (Hockey, 1997). Therefore, more research, especially longitudinal studies addressing long-term effects of imperfect alarm-support on operators' behavioural adaptation, is needed.

A second possible limitation concerns the lack of feedback when participants failed to detect a critical event they were not alerted for by the alarm system. In the present study, critical events were reset automatically if a malfunction remained undiscovered for 10 seconds; no consequences became apparent in this case. However, feedback is critically important for operators to get a clear picture of the level of reliability of a system and to adapt their behaviour accordingly. In a lot of systems, when feedback is not provided or evident, the operator does not know that s/he has failed to detect an alarm system's failure. In real life, misses committed by alarm systems often are linked to severe consequences, albeit these might be delayed somewhat in time (e.g. an overheating of an engine that only after some time leads to a breakdown or engine fire). Nonetheless, in the current study participants still adapted to the alarm systems' reliabilities even without feedback as became evident in the increasing performance in the monitoring task.

3.7 References

Bagheri, N., Jamieson, G. A., 2004. The impact of context-related reliability on automation failure detection and scanning behaviour. In: Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, 212-217.

Bahner, J. E., Hüper, A.-D., Manzey, D., 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies 66, 688-699.

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 in Ergonomics Science 8, 321-348.

Bliss, J. P., Gilson, R. D., Deaton, J. E., 1995. Human probability matching behaviour in response to alarms of varying reliability. Ergonomics 38, 2300-2312.

Bliss, J. P., Dunn, M. C., 2000. Behavioural implications of alarm mistrust as a function of task workload. Ergonomics 43, 1283-1300.

Breznitz, S., 1984. Cry-wolf: The psychology of false alarms. Lawrence Erlbaum Associates, Hillsdale, NJ.

Buchner, A., Erdfelder, E., Faul, F., 1997. How to use G*Power [Computer manual]. Available at: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html.

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. In: Proceedings of the 48th Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA, Human Factors and Ergonomics Society, 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. NASA Langley Research Center, Hampton, VA.

Dixon, S. R., Wickens, C. D., Chang, D., 2004. Unmanned aerial vehicle flight control: False alarms versus misses. In: Proceedings of the 48th Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA, Human Factors and Ergonomics Society, 152-156.

Dixon, S. R., Wickens, C. D., 2006. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. Human Factors 48, 474–486.

Dixon, S. R., Wickens, C. D., McCarley, J. S., 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? Human Factors 49, 564-572.

Getty, D. J., Swets, J. A., Pickett, R. M., Gonthier, D., 1995. System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. Journal of Experimental Psychology: Applied 1, 19-33.

Green, D. M., Swets, J. A., 1966. Signal detection theory and psychophysics. Wiley & Sons, New York.

Hockey, G. R. J., 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. Biological Psychology 45, 73–93.

Lee, J. D., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243-1270.

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

Lees, M. N., Lee, J. D., 2007. The influence of distraction and driving context on driver response to imperfect collision warning systems. Ergonomics 50, 1264-1286.

Manzey, D., Reichenbach, J., Onnasch. L., 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. Journal of Cognitive Engineering and Decision Making 6, 57-87.

Meyer, J., 2001. Effects of warning validity and proximity on responses to warnings. Human Factors 43, 563-572.

Meyer, J., 2004. Conceptual issues in the study of dynamic hazard warnings. Human Factors 46, 196-204.

Meyer, J., Bitan, Y., 2002. Why better operators receive worse warnings. Human Factors 44, 343-354.

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 behaviour. Technical report no. AFRL-RH-WP-TR-2010-0133. Air Force Research Lab, Wright-Patterson, OH (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, 354-365.

Muir, B. M., 1987. Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies 27, 527-539.

Muir, B. M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37, 1905-1922.

Parasuraman, R., Manzey, D., 2010. Complacency and bias in human use of automation: An attentional integration. Human Factors 52, 381-410.

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, 286-297.

Prinzel, L. J., De Vries, H., Freeman, F. G., Mikulka, P., 2001. Examination of automation-induced complacency and individual difference variates. Tech. Memo. no. TM-2001-211413. NASA Langley Research Center, Hampton, VA.

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

Rovira, E., McGarry, K., Parasuraman, R., 2007. Effects of imperfect automation on decision making in a simulated command and control task. Human Factors 49, 76–87.

Sharma, H. O., 1999. Effects of training, automation reliability, personality and arousal on automation-induced complacency in flight simulation task (Ph.D. dissertation). Banaras Hindu University (unpublished).

Singh, I. L., Molloy, R., Parasuraman, R., 1997. Automation-induced monitoring inefficiency: Role of display location. International Journal of Human-Computer Studies 46, 17–30.

Singh, I. L., Sharma, H. O., Singh, A. L., 2005. Effect of training on workload in flight simulation task performance. Journal of the Indian Academy of Applied Psychology 31, 81-90.

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

Swets, J. A. 1992. The science of choosing the right decision threshold in high-stakes diagnostics. American Psychologist 47, 522–532.

Wickens, C. D., Dixon, S. R., Goh, J., Hammer, B., 2005. Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. In: Proceedings of the 13th International Symposium on Aviation Psychology, 919–923. Association of Aviation Psychology, Columbus, OH.

Wickens, C. D., Dixon, S., 2007. The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science 8, 201-212.

Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., Clayton, K., 2009. False alerts in air traffic control conflict alerting systems: Is there a “cry wolf” effect? Human Factors 51, 446-462.

Wiczorek, R., Manzey, D., 2010. Is operators’ compliance with alarm systems a product of rational consideration?, In: Proceedings of the HFES 54th Annual Meeting, Santa Monica, CA, Human Factors Society, 1722-1726.

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, 352-367.

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

4. Study III

Crossing the Boundaries of Automation - Function Allocation and Reliability

Linda Onnasch

Department of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany

Reference

Onnasch, L. (under review). Crossing the Boundaries of Automation - Function Allocation and Reliability. *International Journal of Human-Computer Studies*.

4.1 Abstract

Two important automation characteristics are crucial when considering human performance consequences of automation support. One characteristic concerns the function allocation (FA) between human and automation. Adverse effects of automation seem to be most likely when the human operator is taken out-of-the-loop from active decision-making, excessing a boundary from information automation to decision automation. The second characteristic is the reliability of automation. Previous research suggests a critical reliability boundary around 70% below which automation support cannot be considered as helpful. This study explored differential effects of crossing both boundaries at the same time. Within a multi-task simulation consisting of a monitoring task and two concurrent tasks, participants were assigned to one of six groups, two manual control groups and four automation-supported groups. Automation support differed with respect to FA (stage 1 vs stage 4 automation) and reliability (68.75% vs 87.5%), both factors varied across the critical boundaries. Results suggest that reliability determines human operators' attention allocation and performance. When reliability was below the boundary, participants showed an increased attentional effort and a worse performance compared to fairly reliable support. Against the stated assumptions, FA did not reveal any impact. In combination with previous research this result might indicate that the FA boundary might rather be some kind of "function allocation valley" concerning decision-making automation (stage 3) in which negative consequences for human operators are most likely. Results are discussed in the context of recent automation research.

Keywords: function allocation, stages of automation, reliability, human-automation performance, attention allocation

4.2 Introduction

When considering human-automation interaction, two important automation characteristics are crucial to determine if automation support is beneficial or rather deteriorates performance compared with no automation support. One characteristic concerns the function allocation (FA) between human and automation, i.e. the tasks and functions that are assigned to and consequently carried out by automation. The second characteristic is the reliability of automation that determines to what extent the operator can rely on the proper functioning of the automated system. In recent years, two boundaries have been proposed regarding these two automation characteristics, which are assumed to be critical with respect to human performance consequences. For FA, a number of experimental results comprised in a meta-analysis by Onnasch et al. (2014b) revealed that negative consequences of automation, like loss of situation awareness and manual skills, are most likely when FA moves across a critical boundary from information automation to decision automation (Parasuraman et al., 2000). A second boundary was proposed by Wickens and Dixon (2007) and addresses a critical reliability level of an automated system below which the system cannot be considered beneficial anymore for human performance. Based on a thorough quantitative literature review they provided strong empirical evidence that automation only entails positive effects on joint human-automation performance if the automation's reliability is higher than 0.7 (70%). In case the reliability is lower, working with automation support was found to lead to even worse performance than working without automation support.

The current study builds up on these findings and aims to gain further insight on differential effects when the two previously identified boundaries are both crossed.

4.2.1 Function Allocation between Human and Automation

Different framework models have been proposed to allow a standardised characterisation of automated systems with respect to the distribution of functions between human and automation (Endsley and Kaber, 1999; Endsley and Kiris, 1995; Milgram, et al., 1995; Parasuraman et al., 2000; Riley, 1989; Sheridan, 2000). Common to all these models is the assumption that automation does not exist in an all-or-none fashion, but rather constitutes a continuum from no support to full automation of all functions. Accordingly, potential costs and benefits have to be considered as a function of more or less automation.

A well accepted framework model has been introduced by Parasuraman et al. (2000). They suggest a taxonomy, which characterises automation on two different dimensions: stages and levels. The stage component refers to the human information processing model (Wickens et al., 2013) and differentiates four different functions that can be reallocated to automation: information acquisition (stage 1), information analysis (stage 2), decision-making (stage 3), and action implementation (stage 4). The first two stages are often referred to as information automation (IA), whereas the latter two stages can be summarised by the term decision automation (DA). Additionally, defining the level of automation on each stage further specifies automation. Due to the two-dimensionality of the model, a system can be characterised according to the functions that are automated (stage dimension) and the level of automation on each stage (level dimension). This approach allows comparing various types of automation in a standardised manner and therefore outmatches other models that are only applicable to a certain kind of automation (e.g. Endsley and Kiris, 1995; Milgram, et al., 1995; Riley, 1989; Sheridan, 2000).

Driven by the idea of potential costs and benefits in terms of human performance consequences as a function of more or less automation, the examination of the impact of FA on human-automation interaction has become one of the major interests in automation research. In particular, consequences of different stages of automation on human performance were examined (e.g. Endsley and Kiris, 1995; Kaber et al., 2000; Layton et al., 1994; Lorenz et al., 2002; Manzey et al., 2012). Results suggest that higher stages support human performance optimally by taking over certain parts of the task and thus reducing operators' workload. However, when automation is not perfectly reliable, a higher-stage automation has been shown to adversely affect operators' situation awareness and may also increase the risk of catastrophic failures due to operators' skill loss after a prolonged use of automation. As a consequence, it has been suggested that medium automation would represent the best compromise to maximise performance benefits of automation, and minimise possible new risks at the same time (e.g. Cummings and Mitchell, 2007; Endsley and Kiris, 1995; Kaber et al., 1999; Kaber and Endsley, 1997; Manzey et al., 2012). Applying these results to the automation taxonomy proposed by Parasuraman et al. (2000) the often stated recommendation to implement a mostly not further defined medium automation (Endsley and Kaber, 1999; Endsley and Kiris, 1995; Kaber et al., 1999; Manzey et al., 2012) can be specified. Comparing medium and high amounts of automation in those studies, the most critical difference of these gradations with respect to human performance consequences seems to be whether only information input functions are automated (information automation, stage 1 and 2) or additional output functions like decision-making or action implementation (decision automation, stage 3 and 4).

Direct empirical evidence for this assumption has been provided in a recent meta-analysis by Onnasch et al. (2014b). The meta-analysis was based on 18 experiments and examined the effects of human-automation FA on routine system performance, performance when the automation fails, workload, and situation awareness. Results indicated a clear automation benefit for routine system performance with increasing automation, as well as benefits for workload when automation functions properly. However, when automation does not function properly, i.e. when there is an automation breakdown, a negative impact of more complex automation on failure system performance and situation awareness was reported. This out-of-the-loop-unfamiliarity performance problem (OOTLUF; Wickens, 2000) seems to arise when cognitive functions related to active decision-making are resumed by automation in particular. Accordingly, Onnasch et al. (2014b) found negative consequences to be most likely when automation moves across a critical boundary between stage 2 and stage 3; the latter alleviating the human from active decision-making. When this boundary is crossed the risk of adverse effects on human performance is more likely, as well as potential catastrophic consequences when automation is unreliable or should suddenly fail (Wickens and Hollands, 2000).

4.2.2 Reliability of Automation

The second aspect that is of crucial importance regarding effects of automation on human performance is its reliability (Lee and See, 2004). Depending on the realised stage of automation, reliability can be defined as the proportion of correctly indicated critical events (information automation), correctly given diagnoses, suggested decisions, or correctly executed actions (decision automation).

Several studies have addressed the impact of automation reliability on human performance in terms of complacency (e.g. Parasuraman, et al., 1993), automation bias (for a systematic review see Goddard et al., 2012), situation awareness (e.g. Wickens, 2000), or attention allocation (Wickens et al., 2005). Comprising results of single studies, an overall picture of the impact of automation reliability on human performance was provided by a quantitative literature review conducted by Wickens and Dixon (2007). The analysis included data points from 20 different studies, which explicitly varied the reliability of diagnostic automation. They found a positive linear relation between automation's reliability and the joint human-automation performance. That is, even though operators may have tended to miss more critical events when working with highly reliable automation the overall number of jointly detected critical events was still higher compared to working with less reliable automation. However, when automation reliability was below approximately 70%, automation support yielded even worse performance compared to working with no automation. Thus, effective compensation for unreliability seems to be possible to a certain level only.

However, a drawback of most of the studies reported thus far and that were integrated in Wickens and Dixon's analysis (2007) is that they only compared relatively extreme reliability levels and missed to describe the characteristics of operators' adaptation to automation across a more complete range of reliability. To gain more insights into the proposed change from supportive automation to useless automation Onnasch et al., (2014a) examined the impact of five different reliability levels of alarm systems (the simplest form of information automation) on joint human-automation performance and visual attention allocation in a multi-task simulation. Alarm reliability was set to 68.75%, 75%, 87.5%, or 93.75% by varying the number of critical events that were missed by the alarm system. In comparison with a manual control group they found a clear automation benefit concerning human-automation performance that was independent of the level of automation reliability. Whereas the manual control group only detected around 70% of engine malfunctions, detection rates increased with alarm-support, even in the lowest reliability condition, up to 90%. This result revealed that all participants in the alarm-supported groups adapted to differing reliability levels in a very effective way. However, when reliability was below 70% the performance benefit was associated with an increased attentional effort, and a declined relative performance in a concurrent task compared to the other alarm-supported groups. In fact, when working with the least reliable alarm system, participants allocated as much visual attention to the supported task as the manual control group, i.e. behaved as if no automation support was available. Hence, automation below a reliability of 70% was not beneficial anymore. Similar findings were reported by Wickens et al. (2005) who also found negative consequences for attention allocation strategies and performance when automation reliability was below 70%.

4.2.3 Function Allocation and Automation Reliability

Summarising the scope of research, both, function allocation between human and automation and automation’s reliability, seem to be of critical importance in terms of human performance consequences and adequate, safe human-automation interaction. In particular, when the human operator is taken out-of-the-loop from active decision-making, excessing the boundary from stage 2 to stage 3 automation (Onnasch et al., 2014b), the risk of adverse effects on human performance is most likely. Moreover, research has revealed a critical reliability boundary around 70% below which automation support cannot be considered as helpful anymore (Dixon and Wickens, 2006; Onnasch et al., 2014a; Wickens and Dixon, 2007; Wickens et al., 2005).

However, only few studies have explicitly varied both aspects at the same time (Crocoll and Coury, 1990; Galster and Parasuraman, 2004; Galster, 2003; Rovira et al., 2007; 2002; Sarter and Schroeder, 2001). Some of these studies suggest a sort of interaction effect between FA and reliability. In particular, it was found that same levels of unreliability led to worse effects on human performance in case of decision versus information automation (e.g. Crocoll and Coury, 1990; Rovira et al., 2007; Sarter and Schroeder, 2001). However, other studies report detrimental effects of unreliable automation already for information automation or even worse performance for information compared to decision automation (Galster and Parasuraman, 2004; Galster, 2003).

In the aforementioned studies, the comparison of different reliabilities was often realised with perfect reliability trials compared to completely unreliable trials, in which automation gave only false alarms or missed all events (Crocoll and Coury, 1990; Rovira et al., 2002; Sarter and Schroeder, 2001). Consequently, even fewer studies deliver possible insight into the question of how the critical boundaries related to FA and reliability interact, and still, results are mixed (e.g. Galster and Parasuraman, 2004; Galster, 2003; Rovira et al., 2007). Furthermore, most of the previous research has missed to consider effects of attention allocation induced by different sorts of automation, as well as the joint performance of human and automation.

Therefore, the goal of the current study was to gain further insight into the interaction of FA with the level of reliability and the impact of these automation characteristics on the joint human-automation performance and attention allocation.

The task used for the experiment was a multi-task flight simulation, including three different subtasks. One of these tasks had to be performed with or without automation support. Automation support differed with respect to FA (IA and DA support) and reliability (high and low), with both factors varying across the critical boundaries. The impact of automation on participants' attention allocation was measured via eye-tracking. In addition, performance was assessed for all three tasks. The following hypotheses guided the research.

Attention allocation: Based on the aforementioned results from Onnasch et al. (2014a) and Wickens et al. (2005), it was first of all stated that *support of a fairly reliable automation should relieve participants' attentional demands to the supported task*. Compared to a manual control group, participants should reallocate attention away from the supported task to concurrent tasks. This reallocation induced by reliability should be different for IA and DA support. It was hypothesised that *the attentional relieve should be more substantial for DA support compared to IA* as the complete task is resumed by automation. For unreliable automation support (below 70%) it was expected that participants supported by IA would more or less behave as if the automation support were not available, i.e. would invest as much attention to the automation-supported task as the manual control group (Onnasch et al., 2014a). This should reflect their effort to completely compensate for the automation's imperfection. However, for DA with a low reliability this effect should be less pronounced. Working with DA support should lower participants' willingness to compensate for unreliability because they are taken much more out-of-the loop of task control than with IA support. As a consequence, they may not feel as responsible for the task as with IA because the task is normally almost fully carried out automatically (Dzindolet et al., 2001; Lewandowsky et al., 2000; Mosier et al., 1998; Mosier and Skitka, 1996). Therefore, it was hypothesised that *when reliability undershoots the reliability boundary (70%), effects on attention allocation should be worse for IA than for DA*.

Performance in the automation-supported task: According to results by Onnasch et al. (2014a) it was hypothesised that *independent of reliability, working with IA should result in superior joint human-automation performance compared to no automation support*. Even for the less reliable group, a better overall performance than in the manual control group was expected. This was based on the assumption that participants should realise automation's imperfection and accordingly reallocate their attention to this task. Therefore, besides benefiting from IA support when it correctly indicates malfunctions they should also detect misses of automation which eventually leads to the hypothesised superior performance compared to working manually. *For DA, reliable support should lead to the best joint human-automation performance* compared to all other groups. *However, when reliability falls below the critical boundary, there should be no automation benefit anymore* as operators are out-of-the-loop (Wickens and Hollands, 2000) and do not compensate for automation. This finding would be in accordance to results reported earlier (Crocoll and Coury, 1990; Onnasch et al., 2014a; Rovira et al., 2007; Sarter and Schroeder, 2001).

Performance in concurrent tasks: Based on the finding that more automation is always better when support is reliable (Onnasch et al., 2014b), *the best performance was expected for the group working with a high stage of automation (DA) that does not fall below the critical reliability boundary. The same results, albeit to a weaker extent, were expected for sufficiently reliable IA support and for the group working with DA, even if the reliability of the latter falls below the critical boundary.* The latter assumption is related to the aforementioned hypotheses that participants of this group would not change their attention and performance strategies, even if they realise automation's unreliability. Therefore, they should still maintain a superior performance in concurrent tasks compared to the low reliability IA group, and compared to the manual control group. *Worst concurrent task performance was expected for the low reliability IA group*, as the hypothesised compensation for unreliability should only be possible at the expense of performance in concurrent tasks.

For a better understanding, the three sets of hypotheses are summarised in Figure 1.

Attention to Automated Task Compared to Manual		**Reliability**	
		low	high
Function Allocation	IA	**0**	**-**
	DA	**-**	**--**

Performance Automated Task Compared to Manual		**Reliability**	
		low	high
Function Allocation	IA	**+**	**++**
	DA	**0**	**+++**

Performance Concurrent Tasks Compared to Manual		**Reliability**	
		low	high
Function Allocation	IA	**0 / -**	**+**
	DA	**+**	**++**

Figure 1. Hypotheses related to (1) attention allocation to the automation-supported task (top panel), (2) performance in the automation-supported task (middle panel), and (3) performance in concurrent tasks (lower panel). Symbol meanings are as follows: 0 = no difference compared to the manual control group; - = less attention or worse performance compared to the manual control group; + = more attention or superior performance compared to the manual control group. Two or more symbols (+++) illustrate the strength of the supposed effect compared to the other groups.

Additionally to the scrutinized factors FA and reliability, a time-on-task factor was included in the experimental design as there is evidence that experience with automation may foster or change some of the assumed effects (e.g. adaptation to imperfect automation; Onnasch et al., 2014a).

4.3 Method

4.3.1 Participants

The number of participants was defined based on a power analysis with the assumption of a medium effect size (GPower3.1, for details see e.g. Buchner et al., 1997). This revealed a required sample size of 78 participants. Accordingly, 78 students participated in the experiment. Participants were recruited using a web-based data bank provided by the Institute of Psychology and Ergonomics at the TU Berlin. This tool allows setting certain criteria describing the required sample. Participants for the current study had to be students, aged between 18 and 38, should have German as native language or equal language abilities, had to be frequent computer users, and right hander. These criteria were defined to ensure that participants understood the instructions, and that other factors despite the experimental independent variables did not cause much variance (e.g. educational background).

None of the participants had prior experience with the flight simulation task used in the study. Participation was voluntary and could be cancelled anytime. Participants signed consent forms at the beginning of the experiment and were paid 7€ for completing the study.

Two participants had to be excluded because of problems with the eye-tracker calibration, and four participants because they obviously did not understand the task. Two more students were post-hoc excluded from analyses as their eye-tracking data deviated extremely from those of all other participants[1]. Therefore, data from 70 students (36 female, 34 male) ranging in age from 19 to 36 (M = 25.26, SD = 3.74) were taken into account for the following analyses. Due to the reduced sample size the post-hoc calculated Power for analyses was (1-ß) = .40.

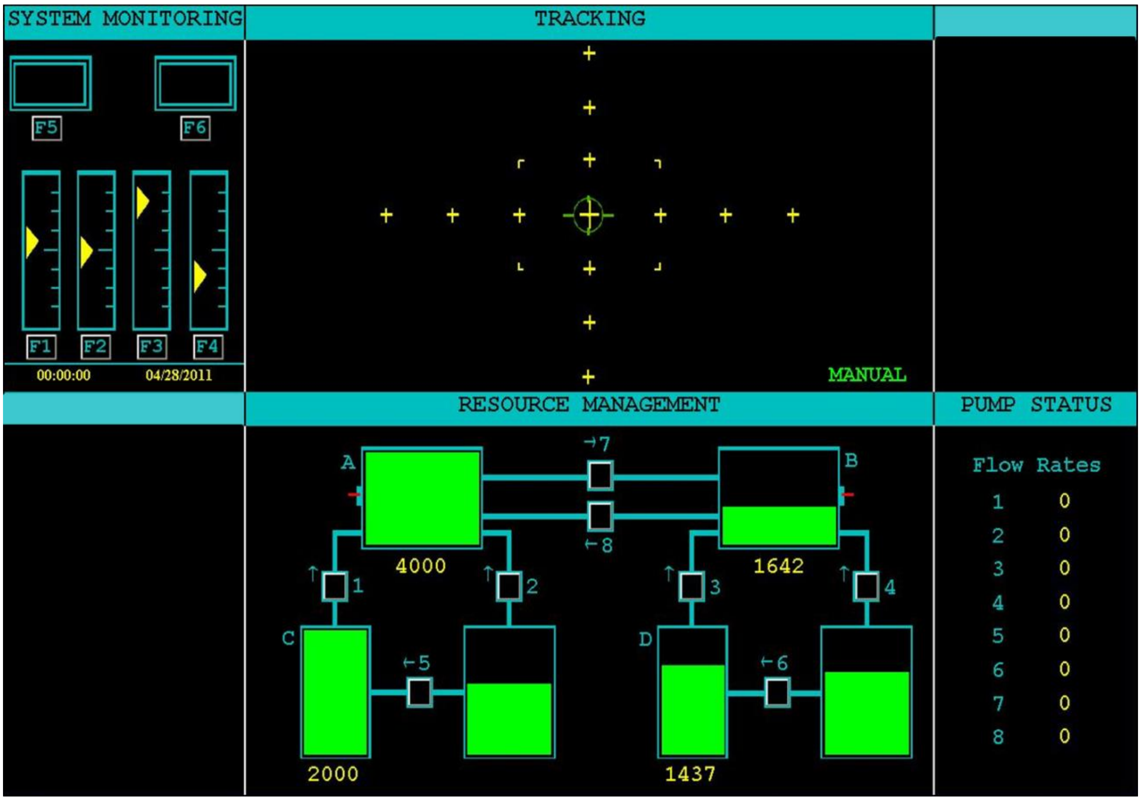
4.3.2 Task and Apparatus

As experimental task the most recent version of the *Multi-Attribute Task Battery* (MATB; Miller, 2010) was used. It was directly based on the original version developed by Comstock and Arnegard (1992) which was used in previous research (e.g. Parasuraman et al., 1993). All main functionalities including the interface corresponded to the original version. Only the programming environment has been changed (MatLab instead of QBasic), which made it easier to implement experimental modifications.

---

[1] The two participants belonged to the group which had a 87% reliable DA support. They were excluded based on an analysis of the dependent variable "mean time between fixations" (MTBF) in the monitoring task. Both participants deviated more than two standard deviations from the mean value, however, in the hypothesised direction. In terms of a conservative testing, both participants were excluded from the entire analyses as the MTBF could also affect other variables like performance in this task or the performance in concurrent tasks.

The MATB is a multi-task flight simulation consisting of a two dimensional compensatory tracking, engine-system monitoring, fuel resource management, communications, and scheduling. In the present study, only the compensatory tracking, the resource management, and the system monitoring were implemented and had to be performed concurrently. The user interface of the MATB used in the present study is shown in Figure 2.



Flow Rates

1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0

Figure 2. MATB as used in the current study with the compensatory tracking in the upper middle position, the resource management beneath and the system monitoring in the upper left display corner.

In the two-dimensional compensatory tracking task participants are required to keep a randomly moving cursor in the centre target position by applying appropriate control inputs via joystick. In the resource management task participants must compensate for fuel depletion by pumping fuel from four supply-tanks into two main tanks.

The system monitoring task was most important for the current research. It consists of four vertical engine gauges with moving pointers that participants must monitor for abnormal values that occur randomly. As long as all engines function properly, the pointers fluctuate by chance within a fixed range around the centre value of the gauges. However, in case of a malfunction the pointer of the gauge for the affected engine suddenly shifts upwards or downwards by two gauge units and starts to fluctuate around this new position. These deviations must be detected by participants and reset by a corresponding key press. If a malfunction is not detected within ten seconds the gauge resets automatically and the event is logged as missed event.

Dependent on task configuration, system monitoring has to be performed manually or with automation support. The first alternative represents a simple binary alarm system (IA support). When working with this system a visual red alert appears above the gauges whenever the automation detects a parameter deviating from its nominal value. Nevertheless, the identification of the affected gauge and the corresponding reset of the parameter still have to be performed manually by the operator. The second alternative consists of DA support. In this case, the visual red alert appears to inform the operator that the automation has detected a deviation on one of the four gauges. Moreover, DA identifies and automatically resets the affected gauge within four seconds. As long as this automation functions properly, the operator does not have to take any evasive action concerning the system monitoring task.

The MATB was presented in front of the participant on a 20 inch monitor that was equipped with a remote eye-tracking system (RED system, SensoryMotoric Instruments, Germany). This system enabled to sample gaze movements during task performance with a sampling rate of 120 Hz. Based on these data, gaze fixations in different areas of interest (AOI, see definition below) were automatically recorded.

4.3.3 Design

The study used a three factorial design with two between-subjects factors (FA and Reliability) and one within-subject factor (Block). The first factor (FA) consisted of two experimental groups. One group was supported in the monitoring task by IA (operationally defined as a binary alarm system), i.e. an automated support clearly below the FA boundary. The other group was supported by DA that resumed the entire task and therefore clearly crossed the FA boundary. The second factor (Reliability) differentiated between two levels of reliability that were chosen with regard to the reliability boundary (Wickens and Dixon, 2007). These reliability levels were already part of a prior experiment and were again chosen in order to allow for a comparison of results (Onnasch et al., 2014a). Furthermore, choosing these specific reliability levels should ensure that groups significantly differed in terms of performance and cognitive demands induced by automation reliability. An impact of reliability was crucial to look for interaction effects of FA and reliability. This implied clearly distinctive reliability levels. Based on these considerations reliability was set to the lowest level that was examined in the prior experiment, 68.75%, and a clearly distinctive level of 87.5% that represented one of the fairly reliable automation conditions in the preceding experiment (Onnasch et al., 2014a). Reliability was realised by varying the number of critical events that were missed by the automation. Additionally, a manual control group was implemented, synonymous with a 0% reliable automation. The third factor (Block) constitutes a within-subject factor and was included to control time-on-task effects that possibly moderate effects of reliability and function allocation. Every participant had to perform the three concurrent tasks of the MATB for three 10-minute blocks. A total of 16 critical events occurred in the monitoring task during each block, which had to be detected by automation or the participant, respectively.

4.3.4 Dependent Measures

To investigate the impact of the experimental factors on the perceived automation's reliability (manipulation check), visual attention allocation and performance, three different categories of dependent measures were sampled and analysed.

Manipulation Check: To assure that participants noticed that the automation did not work perfectly and missed events, participants were asked to rate the *perceived reliability* after they had worked with the automation for a prolonged time. Participants provided ratings to the question "How reliable was the system you worked with" on a visual-analogue scale ranging from 0% to 100%.

Visual attention allocation was measured by means of eye-tracking. Specifically, the *relative fixation time* for different pre-defined areas of interest (AOI) was assessed. For this purpose, three different AOIs (specified by pixel areas) were defined before the experiment started. These AOIs corresponded to the three different tasks participants had to perform: compensatory tracking, resource management, and system monitoring (see Figure 1). Relative fixation time was defined as the time participants fixated an AOI relative to the overall fixation time, i.e. sum of times any one of the AOIs was fixated.

Furthermore, the *mean time between fixations* (MTBF) for the system monitoring AOI was calculated. This variable was defined as the time between the last fixation to the system monitoring AOI and the moment when the AOI was re-entered by a fixation. Therefore, this variable evaluated how much time participants spent on other tasks before they reallocated their attention again to the system monitoring, i.e. how much they relied on the automation to inform them if a system malfunction appears.

Fixations were defined by a minimum duration of 80 ms and a maximum dispersion within this time interval of 100 pixels.

Performance measures were defined for all three tasks of the MATB participants had to perform concurrently and collected for each 10 minute block separately. For the system monitoring task, *percentage of detected system failures* was defined as the percentage of all engine failures detected correctly by the human operator (control condition) or the human and automation together (joint performance in the automation support conditions).

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 calculated as a measure of mean deviation from the central target position, based on deviation data sampled at an interval of 5 seconds. The RMSE for the resource management task was calculated in relation to an optimal tank level, which had to be maintained in both main tanks. Fuel levels were sampled and RMS errors computed for each 5-second period.

4.3.5 Procedure

Following a demographic questionnaire, an instruction on the MATB, and an initial calibration of the eye-tracking system, participants were familiarised with performing the three different tasks manually in a 10 minute practice block. They were instructed that all three tasks would be of equal importance, and that they should work on all tasks concurrently with equal priority. Afterwards, they were randomly assigned to one of the six groups. Participants in the four experimental groups were introduced to the automation they were subsequently working with. Specifically, they were told which functions were resumed by automation. Furthermore, they were informed, that the automation would not be perfectly reliable so that participants may not fully rely on it. However, no concrete reliability information was provided. Then, the experiment started consisting of three 10 minute blocks. Prior to each block the eye tracker was re-calibrated. The perceived reliability of the alarm system was assessed in the experimental groups after the second block. The experiment ended with the debriefing of participants.

4.4 Results

In a first step, eye-tracking and performance measures of the manual control groups were analysed to make sure that the two groups did not differ and could serve as comparable baselines for the experimental groups. Because no differences were expected, α was adapted to a 20% level for these analyses (null-hypothesis testing). The 2 (Control Groups) x 3 (Block) ANOVA's did not reveal any significant differences between the two control groups.

4.4.1 Perceived Reliability

A two-factorial 2 (FA) x 2 (Reliability) between-subjects ANOVA contrasted the four experimental groups with automation support that were considered for this analysis. Results revealed a significant difference in the perceived reliability ratings contrasting the groups with 68.75% and 87.5% reliable automation support, $F(1, 42) = 13.11$, $p < .002$, $\eta^2 = .23$. The low reliability group rated their automation support to be 68.20% dependable whereas the high reliability group estimated the reliability of their automation support to be 81.97%. No differences were found for the factor FA or an interaction of the two factors (both $F < 1.0$). Results confirm that the manipulation was successful, as perceived reliabilities were systematically related to the actual ones and significantly differed between experimental reliability conditions.

4.4.2 Attention Allocation

Figure 3 illustrates the results for the *relative fixation time* on the three different AOIs, i.e. monitoring task (left panel), tracking task (middle panel) and resource management task (right panel). The upper part of the Figure represents participants' results working with information automation (IA), the lower part shows results for the decision automation groups (DA).

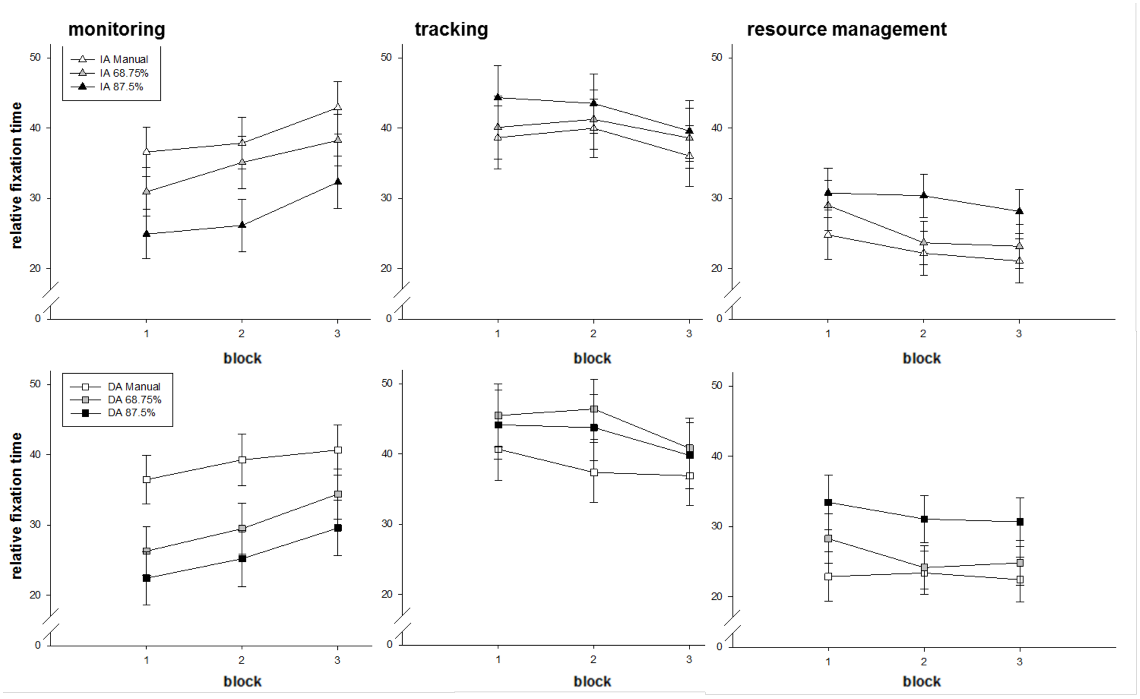


Figure 3. Effects of reliability and function allocation (upper panels IA support, lower panels DA support) on the relative fixation time; AOI from left to right: monitoring, tracking and resource management.

For the monitoring task, increments of reliability led participants to spend less visual attention on the monitoring task. As expected, participants in the manual control groups had the longest relative fixation time to this AOI ($M_{manual} = 38.95\%$), followed by groups working with automation that violated the critical reliability boundary ($M_{68.75\%} = 32.40\%$). Participants working with 87.5% reliable automation fixated this AOI less often ($M_{87.5\%} = 26.75\%$) and reallocated their visual attention more to the concurrent tasks. Furthermore, with ongoing time-on-task, there was a tendency in all groups to allocate more attention to the monitoring AOI. A 2 (FA) x 3 (Reliability) x 3 (Block) ANOVA supported this pattern of results statistically with a main effect of Reliability, $F(2, 64) = 6.34$, $p < .004$, $\eta^2 = .16$, and a main effect of Block, $F(1.65, 105.74) = 29.3$, $p < .001$, $\eta^2 = .31$. Interestingly, post-hoc analyses revealed that the pattern of visual attention allocation of participants with the 68.75% automation support was comparable to working without automation ($p_{manual\text{-}68.75\%} = .15$). In contrast, automation that was fairly reliable led to a reallocation of attention away from the automation-supported task ($p_{manual\text{-}87.5\%} < .004$). All effects were independent of Function Allocation ($F < 1$).

For visual attention allocation to the tracking task, only a main effect of Block reached significance as all participants reduced fixation times to this AOI through blocks, $F(2, 128) = 7.10$, $p < .002$, $\eta^2 = .10$.

As can be seen in Figure 3 (right panel) attention allocation strategies to the resource management differed with respect to reliability. The groups working with the 87.5% reliable support had the longest fixation times, followed by participants working with 68.75% reliable support and manual control groups. This was statistically supported by a main effect of Reliability, $F(2, 64) = 3.21$, $p < .05$, $\eta^2 = .09$. Furthermore, all groups spent more time on this AOI in the first block than in the second or third, $F(2, 128) = 9.55$, $p < .001$, $\eta^2 = .13$.

Post-hoc tests revealed that only the support of automation that exceeded the reliability boundary led to an attentional reallocation to this concurrent task that was (marginally) different from working completely manually ($p_{\text{manual-87.5\%}} = .05$). No differences to the manual control group were found for groups working with automation support that was less reliable than 70% ($p_{\text{manual-68.75\%}} = .68$).

The *MTBF* was calculated for the system monitoring AOI. As can be seen in Figure 4 the MTBF was highly dependent on automation's reliability. Participants working with the 87.5% reliable automation had the longest interim time not looking to the automation-supported task, followed by participants supported by a less reliable automation and the manual control groups. On a descriptive level, it seems that also the FA has a certain impact on MTBF as the DA experimental groups differ from their counterparts in the IA groups by having longer MTBF. However, analyses only revealed a main effect of Reliability, $F(2, 64) = 5.05$, $p < .01$, $\eta^2 = .13$, but no interaction of Reliability and FA ($F< 1$). As for the relative fixation time, post-hoc tests revealed that only the MTBF of the 87.5% condition significantly differed from the manual control group ($p = .01$) but not the MTBF of the 68.75% condition ($p = .29$). Furthermore, there was a main effect of Block, $F(2, 128) = 22.17$, $p < .001$, $\eta^2 = .25$, as MTBF decreased with time-on-task (B1 = 3.80s, B2 = 3.53s, B3 = 2.83s).

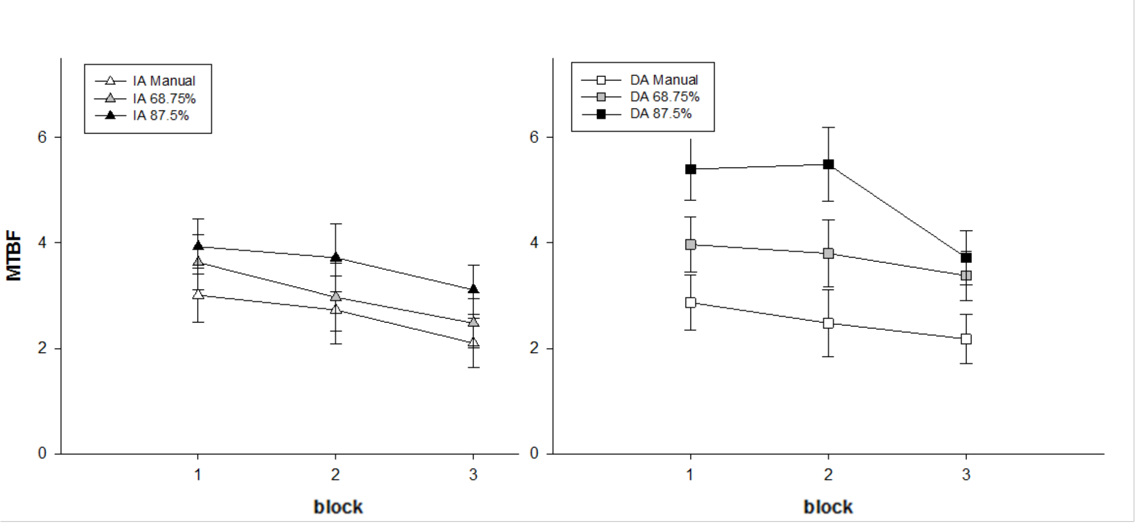


Figure 4. Effects of reliability and function allocation (left panel IA support, right panel DA support) on the MTBF for the monitoring AOI.

4.4.3 Performance Measures

4.4.3.1 Monitoring Task

Results for the *percentage of detected system failures* are illustrated in Figure 5. Performance improved for all groups with time-on-task. Best results were reached by groups that were supported by automation with a reliability above the critical boundary ($M_{87.5\%} = 94.53\%$), followed by participants supported by automation undershooting the critical reliability boundary ($M_{68.75\%} = 88.19\%$) and participants with no automation support ($M_{manual} = 81.42\%$). Statistically, this pattern of results was mirrored in a significant main effect of Block, $F(2, 128) = 12.83$, $p < .001$, $\eta^2 = .16$, and a main effect of Reliability, $F(2, 64) = 9.25$, $p < .001$, $\eta^2 = .22$. Post-hoc tests revealed that only groups supported by a 87.5% reliable automation benefited from their support as they showed superior performance compared to the manual control group ($p_{manual\text{-}87.5\%} < .001$). Performance of participants working with a 68.5% reliable automation support did not significantly differ from performance of the control group which had to detect all system malfunctions manually ($p_{manual\text{-}68.75\%} = .08$). No effects were found for Function Allocation ($F < 1$).

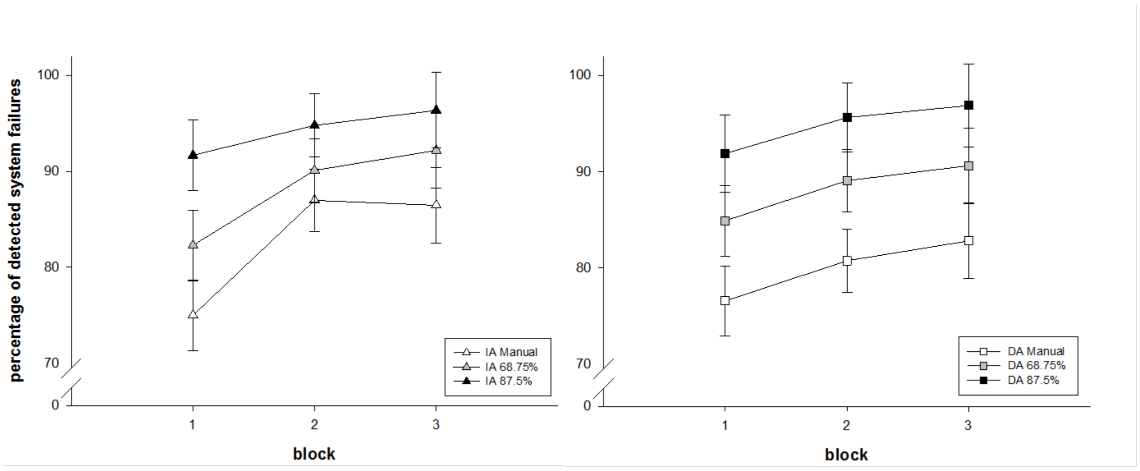


Figure 5. Effects of reliability and function allocation (left panel IA support, right panel DA support) on the overall percentage of detected system failures.

4.4.3.2 Concurrent Tasks

Analyses of RMSE in concurrent tasks only revealed time-on-task effects, F_{Tracking} (1.58, 101.14) = 4.40, $p < .06$, $\eta^2 = .06$; $F_{\text{Resource Management}}$ (1.50, 96.46) = 16.83, $p < .001$, $\eta^2 = .20$. Independent of experimental condition, all groups showed reduced deviations from a given optimal level with increasing experience in the tracking task ($B_1 = 147.60$, $B_2 = 143.33$, $B_3 = 140.78$) as well as in the management of fuel resources ($B_1 = 152.33$, $B_2 = 131.01$, $B_3 = 121.23$).

4.5 Discussion

The study's main objective was to investigate the interaction of FA with the level of reliability and its impact on human operators' attention allocation and the joint human-automation performance.

Therefore, three sets of hypotheses were stated, which related to participants' attention allocation, the joint human-automation performance in the supported task as well as participants' performance in concurrent tasks. An overall contrast of hypotheses and results is presented in Figure 6, which points out that the hypotheses were only partially supported by the data.

Hypotheses

Attention to Automated Task Compared to Manual		Reliability	
		low	high
Function Allocation	IA	0	-
	DA	-	--

Performance Automated Task Compared to Manual		Reliability	
		low	high
Function Allocation	IA	+	++
	DA	0	+++

Performance Concurrent Tasks Compared to Manual		Reliability	
		low	high
Function Allocation	IA	0 / -	+
	DA	+	++

Results

Attention to Automated Task Compared to Manual		Reliability	
		low	high
Function Allocation	IA	0	-
	DA	0	-

Performance Automated Task Compared to Manual		Reliability	
		low	high
Function Allocation	IA	0	+
	DA	0	+

Performance Concurrent Tasks Compared to Manual		Reliability	
		low	high
Function Allocation	IA	0	0
	DA	0	0

Figure 6. Hypotheses (left side) and actual results (right side) related to (1) attention allocation to the automation-supported task (top panel), (2) performance in the automation-supported task (middle panel), and (3) performance in concurrent tasks (lower panel). Symbol meanings are as follows: 0 = no difference compared to the manual control group; - = less attention or worse performance compared to the manual control group; + = more attention or superior performance compared to the manual control group. Two or more symbols (+++) illustrate the strength of the supposed effect compared to the other groups.

Results for attention allocation revealed a strong impact of automation reliability on participants' attentional strategies. As expected, participants who were supported by 87.5% reliable automation reallocated attention away from the supported task to concurrent tasks. Therefore, fairly reliable automation relieved participants' attentional demands to this task. In contrast, automation that was less reliable (68.75%) was not beneficial as participants invested as much attention to the automation-supported task as participants working manually. However, function allocation did not seem to have an impact on this effect, as no differences could be found between groups that were supported by IA and groups with DA. Crossing both boundaries did not reveal differential effects compared to crossing just the reliability boundary (Wickens and Dixon, 2007). Benefits of fairly reliable support and costs for unreliable automation were the same for IA and DA. Therefore, "more automation" was not "always better" (see also Onnasch et al., 2014b) but fairly reliable automation was indeed always better than unreliable automation.

Effects for performance measures mirrored the effects found for attention allocation. Performance in the automation-supported task revealed clear benefits of automation above the critical reliability boundary compared to working without automation.

However, when supported by automation less reliable than 70%, no automation benefit in terms of a superior performance could be observed. In combination with results for attention allocation, this finding is of special interest as it underlines the severity of undershooting the reliability boundary. Even though participants showed more attentional effort to the automation-supported task than the 87.5% reliability groups, they could not compensate for automation's unreliability. Furthermore, even with a detection rate of 68.75% by automation, the joint human-automation performance did not exceed performance of completely unsupported performance on this task. In short, automation support with reliability below the boundary was not supportive at all! This finding emerged completely independent of the function allocation between human and automation. Highly reliable DA support did not facilitate a superior performance compared to the IA counterpart and low reliable DA support did not reveal worse consequences compared to the IA 68.75% condition.

The overall pattern of results did not change when performance in concurrent tasks was additionally considered. Regardless of the concurrent task, results did not further differentiate findings as only time-on-task effects emerged but no effects of function allocation or automation's reliability.

In summary, the current study revealed two important findings. First, automation reliability determines human operators' attention allocation and performance. When reliability is relatively low, particularly below 70% as stated by Wickens and Dixon (2007), operators invest as much visual attention to the supported task as operators working completely manually and performance does not benefit either. These results are even stronger than expected based on a prior study that showed detrimental effects only on the attentional level (Onnasch et al., 2014a). The attentional effort participants invested when supported by 68.75% reliable automation is comparable in both studies. However, whereas in the first experiment (Onnasch et al., 2014a), more attentional effort enabled participants to protect performance in the supported task, the attentional compensation for automation's imperfection did not prove successful in the current study.

The second important finding is related to the joint impact of reliability and function allocation. Contrary to the assumptions, FA did not reveal any impact on attention allocation or performance. This finding is in contrast to our prior meta-analysis that has provided clear evidence for impacts of function allocation on human performance (Onnasch et al., 2014b). It is also in contrast to the few studies that could provide evidence for a more severe effect of unreliability in interaction with higher-stage than lower-stage automation (Crocoll and Coury, 1990; Rovira et al., 2007; Sarter and Schroeder, 2001)

In those studies higher-stage automation was realised as stage 3 automation (decision support). The current experiment implemented automation that not only resumed decision-making but also the action implementation component of the monitoring task (stage 4). This difference could be crucial for operators perceiving their role and accountability in interaction with automation. When a task is completely automated, the operator's role is that of a supervisor, i.e. he has to monitor the proper functioning of the automation and ensure that everything is as expected. The human operator as a supervisor is accountable for the overall task. In contrast to that, the role definition of an operator in interaction with stage 3 automation might not be that clear-cut. Whereas the human operator is taken out-of-the-loop from cognitive processes of a task, he is still responsible for the implementation of an action that was supposed by automation. Therefore, an operator who is still in charge of action implementation may define his role more or less limited to this subtask instead of feeling responsible for the entire task.

These differently perceived role images might explain the deviant findings between prior research and the current results. Based on the foregoing reasoning it seems plausible that operators supported by stage 3 automation stay mainly focussed on executing what the automation recommends. This might reduce their probability to detect an automation's malfunction and make them more prone to adverse performance effects resulting from unreliability of automation. This fits to the findings of previous research. In contrast, operators who are supported by automation that resumes the complete task, as in the current experiment, are focussed on monitoring the automation's proper functioning, and therefore should be able to compensate for it at least partially if a malfunction occurs. This role perception as a supervisor of automation therefore resulted in comparable attention allocation strategies to the IA-supported group which had to allocate attention to the automation-supported task because those participants were still accountable for the concrete task fulfilment. Therefore, although the role images behind the expressed behaviour of IA- and DA-supported groups might differ, both role perceptions resulted in the same behavioural outcome.

Support for this interpretation is provided by Lorenz et al. (2002). They compared participants' information sampling and performance when working in a multi-task simulation. In one out of three tasks participants were either supported by automation that helped them with information acquisition (stage 1), with decision-making by providing diagnoses of system state and according actions (stage 3) or they were supported by automation that resumed the complete task (stage 4). Results indicated that the group supported by stage 3 automation significantly reduced automation monitoring and showed worse performance when resuming the task from automation compared to the other two groups. Lorenz et al. (2002) suppose that participants with decision-making support may have taken “the displayed advisory [of automation] more as a directive and subsequently focused on its implementation” (p. 895). In contrast to that, the monitoring behaviour and performance of participants supported by full automation (stage 4) did not differ from participants that were only supported by information acquisition automation (stage 1). Lorenz et al. (2002) assert that by verifying automation’s proper functioning participants supported by full automation (stage 4) developed better system knowledge compared to participants with decision-making automation (stage 3), and therefore could appropriately resume tasks from automation under conditions of automation breakdown.

In the current study, the tendency to verify automation’s proper functioning could have been additionally boosted by the fact, that the verification was easily possible. Participants had direct access to relevant data as the monitoring window was visible at all times. There was no extra effort that had to be invested to access raw data. However, accessibility of information has an impact on participants’ verification behaviour of automation. When verification is easy, operators tend to cross-check diagnoses more often than if accessibility of information is more costly, e.g. time consuming. This tendency is still observable even if this behaviour does not represent the best strategy, e.g. in terms of possible gains (Manzey et al., 2014).

With regard to practical implications, results first suggest that automation that hurts the reliability boundary proposed by Wickens and Dixon (2007) should not be implemented, as operators do not benefit from this support. Even though participants tried to adapt to unreliability, the higher attentional effort compared with operators of higher reliable systems did not lead to a superior performance compared to working manually on the task. Concerning the impact of function allocation prior research has suggested that automation support of decision-making functions and beyond (stage 3 and higher) might be particularly critical as negative consequences for human operators were observed with stage 3 automation which did not occur with lower-stage automation. The results of the present study qualify this conclusion by showing that this might be a specific effect of stage 3 automation but may not be true for automation that goes beyond. If the entire task is automated, including action implementation, operators might perceive their role more as a supervisor of automation. Although they are completely taken out-of-the-loop in terms of task completion, they are to a lesser extent out-of-the-loop in terms of task understanding. Therefore, the boundary of function allocation might not be a real boundary but rather some kind of “function allocation valley” concerning stage 3 automation in which negative consequences for human operators are most likely.

4.6 References

Bahner, J. E., Hüper, A.-D., Manzey, D., 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies 66, 688–699.

Buchner, A., Erdfelder, E., Faul, F., 1997. How to use G*Power [Computer manual]. Available at: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html.

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

Crocoll, W. M., Coury, B. G., 1990. Status or recommendation: Selecting the type of information for decision aiding. In: Proceedings of the 34th Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA, Human Factors and Ergonomics Society, 1524–1528.

Cummings, M. L., Mitchell, P. J., 2007. Operator scheduling strategies in supervisory control of multiple UAVs. Aerospace Science and Technology 11, 339–348.

Dixon, S. R., Wickens, C. D., 2006. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. Human Factors 48, 474–486.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., 2002. The perceived utility of human and automated aids in a visual detection task. Human Factors 44, 79-94.

Endsley, M. R. 1987. The application of human factors to the development of expert systems for advanced cockpits. In: Proceedings of the 31st Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA, Human Factors and Ergonomics Society, 1388–1392.

Endsley, M. R., Kaber, D. B. 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42, 462–492.

Endsley, M. R., Kiris, E. O. 1995. The out-of-the-loop performance problem and level of control in automation. Human Factors 37, 381–394.

Galster, S. M. 2003. An examination of complex human-machine system performance under multiple levels and stages of automation. Technical memorandum no. AFRL-HE-WP-TR-2003-0149. Wright-Patterson Air Force Base, Air Force Research Laboratory, OH.

Galster, S. M., Parasuraman, R., 2004. Task dependencies in stage-based examinations of the effects of unreliable automation. In: Proceeding of the Second Human Performance, Situation Awareness and Automation Conference, 23–27.

Goddard, K., Roudsari, A., Wyatt, J. C., 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19, 121–127.

Kaber, D. B., Endsley, M. R., 1997. Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. Process Safety Progress 16, 126–131.

Kaber, D. B., Omal, E., Endsley, M., 1999. Level of automation effects on telerobot performance and human operator situation awareness and subjective workload. Automation technology and human performance: Current research and trends, 165–170.

Kaber, D. B., Onal, E., Endsley, M. R., 2000. Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. Human Factors and Ergonomics in Manufacturing & Service Industries 10, 409–430.

Layton, C., Smith, P. J., McCoy, C. E., 1994. Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation. Human Factors 36, 94–119.

Lee, J., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243–1270.

Lewandowsky, S., Mundy, M., Tan, G., 2000. The dynamics of trust: Comparing humans to automation. Journal of Experimental Psychology: Applied 6, 104–123.

Lorenz, B., DiNocera, F., Röttger, S., Parasuraman, R., 2002. Automated fault-management in a simulated spaceflight micro-world. Aviation, Space, and Environmental Medicine 73, 886–897.

Madhavan, P., Wiegmann, D. A., Lacson, F. C., 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. Human Factors 48, 241–256.

Manzey, D., Gérard, N., Wiczorek, R., 2014. Decision-making and response strategies in interaction with alarms: The impact of alarm reliability, availability of alarm validity information, and workload. Ergonomics (ahead-of-print).

Manzey, D., Reichenbach, J., Onnasch, L., 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. Journal of Cognitive Engineering and Decision Making 6, 57–87.

Milgram, P., Rastogi, A., Grodski, J. J., 1995. Telerobotic control using augmented reality. In: 4th IEEE International Workshop on Robot and Human Communication, 21–29.

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 behaviour. Technical report no. AFRL-RH-WP-TR-2010-0133. Air Force Research Lab, Wright-Patterson, OH
(Retrieved from http://dodreports.com/pdf/ada537547.pdf).

Mosier, K. L., Skitka, L. J., 1996. Human decision makers and automated decision aids: Made for each other? Automation and human performance: Theory and applications, 201–220.

Mosier, K. L., Skitka, L. J., Heers, S., Burdick, M., 1998. Automation bias: Decision making and performance in high-tech cockpits. The International Journal of Aviation Psychology 8, 47–63.

Onnasch, L., Ruff, S., Manzey, D., 2014a. Operators' adaptation to imperfect automation – impact of miss-prone alarm systems on attention allocation and performance. International Journal of Human-Computer Studies 72, 772–782.

Onnasch, L., Wickens, C. D., Li, H., Manzey, D., 2014b. Human performance consequences of stages and levels of automation: An integrated meta-analysis. Human Factors 56, 476–488.

Parasuraman, R., Molloy, R., Singh, I. L., 1993. Performance consequences of automation-induced „complacency". The International Journal of Aviation Psychology 3, 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, 286–297.

Riley, V., 1989. A general model of mixed-initiative human-machine systems. In: Proceedings of the 33rd Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA, Human Factors and Ergonomics Society, 124–128.

Rovira, E., McGarry, K., Parasuraman, R., 2007. Effects of imperfect automation on decision making in a simulated command and control task. Human Factors 49, 76–87.

Rovira, E., Zinni, M., Parasuraman, R., 2002. Effects of information and decision automation on multi-task performance. In: Proceedings of the 46th Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA, Human Factors and Ergonomics Society, 327–331.

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, J. V., Guilfoos, B. J., 2004. Exploring automation issues in supervisory control of multiple UAVs. In: Proceedings of the Human Performance, Situation Awareness, and Automation Technology Conference, 218-222.

Sarter, N. B., Schroeder, B., 2001. Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. Human Factors 43, 573–583.

Sheridan, T. B., 2000. Function allocation: Algorithm, alchemy or apostasy? International Journal of Human-Computer Studies 52, 203–216.

Sheridan, T. B., Verplank, W. L., 1978. Human and computer control of undersea teleoperators. Technical report. MIT Man-Machine Systems Laboratory, Cambridge, MA.

Wickens, C. D., Hollands, J. G., 2000. Engineering psychology and human performance. Prentice Hall, New York, NY.

Wickens, C. D., 2000. Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness. Technical report no. ARL-00-10/NASA-00-2. NASA Ames Research Center, Moffett Field, CA.

Wickens, C. D., Dixon, S. R., 2007. The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science 8, 201–212.

Wickens, C. D., Dixon, S. R., Goh, J., Hammer, B., 2005. Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. In: Proceedings of the 13th International Symposium on Aviation Psychology, 919–923. Association of Aviation Psychology, Columbus, OH.

Wickens, C. D., Hollands, J., Banbury, S., Parasuraman, R., 2013. Engineering Psychology & Human Performance (4th edition.). Pearson, Boston.

5. General Discussion

The objective of this thesis was to gain further insight into the effects of function allocation and automation reliability on human operators' performance and cognitive demands. For this purpose, Parasuraman, Sheridan and Wickens' flow chart model for automation design (2000) served as a guiding framework. The model attempts to help developers to choose an appropriate function allocation between human and automation. For that reason, the model's consecutive structure serialises relevant aspects that have to be considered in design. After a first consideration of what should be automated the model suggests primary criteria to evaluate possible consequences of a proposed function allocation in terms of human performance consequences and imposed cognitive demands. When the initial function allocation has held out against this primary evaluation, the secondary evaluative criterion of reliability has to be considered to assure an appropriate decision regarding the final function allocation. In line with the model's consecutive structure three studies were conducted. The results of the single studies have already been discussed in the according chapters. Therefore, this chapter provides a recapitulation of results (chapter 5.1) to introduce the main objective of the general discussion: the relation of findings to the flow chart model proposed by Parasuraman et al. (2000; chapter 5.2). Furthermore, the three conducted studies are subject to some critical considerations concerning the applied methodical approaches in the single studies in particular (chapter 5.3). The thesis concludes with an outlook, in which future research opportunities are discussed and results are related to trends in human-automation interaction like adaptive automation (chapter 5.4).

5.1 Summary of Results

The first study presented in this thesis is a meta-analysis, which addressed impacts of different function allocations on Parasuraman et al.'s primary evaluative criteria: operators' performance and cognitive demands (Parasuraman et al., 2000). The meta-analysis included 18 experiments that allowed examining effects of function allocation on routine system performance, performance when the automation failed (automation breakdown), workload, and situation awareness.

Results of this meta-analysis revealed a cost-benefit trade-off with regard to function allocation. Under normal operating conditions automation supports users in terms of performance in the automation-supported task. This benefit is positively related to the degree of automation, i.e. the higher the degree of automation, the higher the joint human-automation performance. However, under conditions of automation breakdown, this relation is inversed, i.e. an increasing degree of automation leads to worse performance when the operator has to resume the task. Therefore, routine and failure aspects of performance have contrary effects with increasing automation. This trade-off was also found for workload and situation awareness. Whereas higher degrees of automation reduce operators' workload, the risk of a loss of situation awareness increases. Consequently, an appropriate function allocation can only serve two of the four examined aspects: performance under routine conditions and workload on the one hand, or performance under conditions of automation breakdown and situation awareness on the other hand.

A post-hoc analysis specified this finding as benefits and costs of more or less automation were related to the stages dimension of automation, i.e. what stages of human information processing are supported by an automated system (Parasuraman et al., 2000). In this regard, the differentiation of information automation and decision automation is crucial in determining costs and benefits of automation support. Raising automation from information acquisition and analysis to a support of decision-making makes the human operator much more vulnerable to a loss of situation awareness and manual skills. At the same time, performance under routine conditions, as well as operators' workload, considerably benefit from this extension of automated functions.

In the context of the flow chart model (Parasuraman et al, 2000) these results provide specific insight into effects of function allocation on the primary evaluative criteria that has not been provided before. This specifies the model as the findings of the meta-analysis allow not only stating which aspects should be considered in automation design but also providing concrete advices in relation to the stage component of automation.

To further specify the flow chart model, the second study presented in this thesis focussed on an experimental investigation of the impact of the most important secondary evaluative criterion, i.e. an automation's reliability, on human performance consequences. More specifically, the objective of this study was to provide further insight into effects of different levels of reliability on operators' adaptation in terms of overall system performance as well as operators' cognitive demands. Within a multi-task simulation consisting of a monitoring task and two concurrent tasks, participants were assigned to one of five groups. In the manual control group none of the tasks were supported by automation, whereas the four experimental groups were supported in the monitoring task by a miss-prone alarm system (information automation) differing in reliability: 68.75%, 75%, 87.5%, and 93.75%, respectively. These reliability levels were chosen to allow a comparison of operators' adaptation to a fine grained range of reliability levels instead of comparing only two extreme reliabilities as was done in most past research (e.g. Dixon & Wickens, 2006; Dixon, Wickens, & Chang, 2004; Dixon, Wickens, & McCarley, 2007; Parasuraman, Molloy, & Singh, 1993; Wickens, Dixon, Goh, & Hammer, 2005). Moreover, particularly the lower reliability levels were realised to further investigate a proposed reliability boundary around 70%. This boundary was first stated by Wickens and Dixon (2007) based on findings of a meta-analysis regarding benefits of imperfect automation. The reliability boundary states that the joint human-automation performance becomes even worse than when working manually on a task when automation support is less reliable than 70%. In other words, automation undershooting this boundary does in fact not support task performance but rather deteriorates it.

The study revealed clear automation benefits in the automation-supported task compared to working manually on the task. No differences emerged between groups working with different reliability levels. Therefore, results suggest that participants adapted to differing reliability levels in a very effective way. Costs of unreliable support, particularly automation undershooting the proposed reliability boundary of 70% (Wickens & Dixon, 2007), became evident in concurrent task performance and in attention allocation strategies. With increasing experience, participants supported by automation less reliable than 70% started to change their behaviour according to automation's imperfection by re-allocating their attention away from one of the concurrent tasks. The attentional shift was so pronounced that the adapted attention strategy was comparable to the group working with no automation support. By applying this strategy, operators could compensate for the improper functioning of automation, even for this level of unreliability. Yet compensation was only feasible at the expense of a highly demanding attention allocation strategy, which eventually led to a relative performance decline in concurrent task performance.

When reliability was above 70% participants benefited from automation not only in terms of task performance but also by relieving participants' cognitive demands to the automation-supported task. This was particularly true for highly reliable automation, i.e. cognitive demands decreased with increasing reliability.

Based on findings of the first and second study, the main objective of the lastly presented experiment was to investigate the interaction of function allocation with the level of reliability and its impact on human operators' cognitive demands in terms of attention allocation, and overall system performance. Within the same multi-task simulation as in the second study participants were assigned to one of six groups, two manual control groups and four automation-supported groups. Automation support differed with respect to function allocation and reliability. Both factors were varied across the proposed critical boundaries: regarding function allocation, automation was either realised as information automation or decision automation, thus crossing the boundary between stage 2 and stage 3 (Onnasch, Wickens, Li, & Manzey, 2014). Reliability was set to 68.75% and 87.00%, respectively, thus crossing the boundary of 70% (Onnasch, Ruff, & Manzey, 2014; Wickens et al., 2005; Wickens & Dixon, 2007).

Results for attention allocation again revealed a strong impact of automation reliability on participants' attentional strategies. A fairly reliable automation (above 70%) relieved participants' attentional demands. In contrast, an automation that violated the reliability boundary was not beneficial, as participants invested as much attention to the automation-supported task as participants working manually. Furthermore, the joint human-automation performance in the supported task did not exceed completely unsupported performance although participants showed more attentional effort to the supported task than groups working with a more reliable automation. This pattern of results underlines the severity of undershooting the reliability boundary. In short, automation support that is less reliabel than 70% is not supportive at all!

However, function allocation did not seem to have any impact on the reported effect. Crossing both boundaries (70% reliability boundary and boundary between stage 2 and stage 3 automation) did not reveal differential effects compared with crossing just the reliability boundary. As described in the discussion of this experiment in some more detail (chapter 4.5), a possible explanation for this non-finding relates to the realisation of the high-stage automation which was different to that used in other research. Whereas prior studies have compared information automation with decision-making automation (e.g. Rovira, McGarry, & Parasuraman, 2007) this study implemented high-stage automation by automating the entire task. This changed the operators' role to that of a supervisor by taking the human completely out-of-the-loop. However, instead of increasing risks that are associated with increasing automation this might have taken the human in-the-loop again in terms of task understanding. As a supervisor participants could have felt responsible for the entire task, which also incorporated monitoring the automation's proper functioning. Accordingly, results revealed that participants monitored the automated task as frequent as participants who were supported only by a basic automation (IA).

These results, as well as the findings of the latter two studies, revealed important implications for automation design. Therefore, the next chapter aims at relating those findings to the flow chart model (Parasuraman et al., 2000) to specify the evaluation process of automation and to support designers with concrete guidelines.

5.2 Specification of the Flow Chart Model

The claim of the flow chart model by Parasuraman et al. (2000) is to provide a simple guideline for decisions of function allocation. The model can be used by designers as a starting point for considering what functions should be automated to what extent. Moreover, the model can be understood as a framework for research in the automation domain as it raises (hypothesised) relevant issues that might impact intended benefits of automation implementation. Furthermore, the integration of the two-dimensional model of automation characteristics (stages and levels) allows an attribution of benefits and costs of automation to a very detailed level. Additionally, the structuring of the model as a flow chart supports developers of automation as this defines which aspects should be considered first in evaluating an initial function allocation and which factors should be focussed on in later steps. This outmatches other models that are rather descriptive and often only provide a tangled mass of aspects that should be somehow considered instead of providing guidance in function allocation (e.g. Endsley & Kiris, 1995; Milgram, Rastogi, & Grodski, 1995; Riley, 1989, 1996; Sheridan, 2000).

However, the model's promising structure is not further explicated with regard to concrete contents. Besides the notion of which aspects should be overall considered in automation design, the model in the original form only provides a hypothetical example of function allocation but no tangible guidelines. Therefore, results of the three reported studies are associated to the flow chart model to specify the evaluation process of automation and to support designers with concrete guidelines. Specifications as well as extensions of the model are shown in figure 1 (highlighted in blue). These are described in detail in the following, beginning from the top (guiding questions) to the bottom (final automation) of the model with respect to the conducted studies.

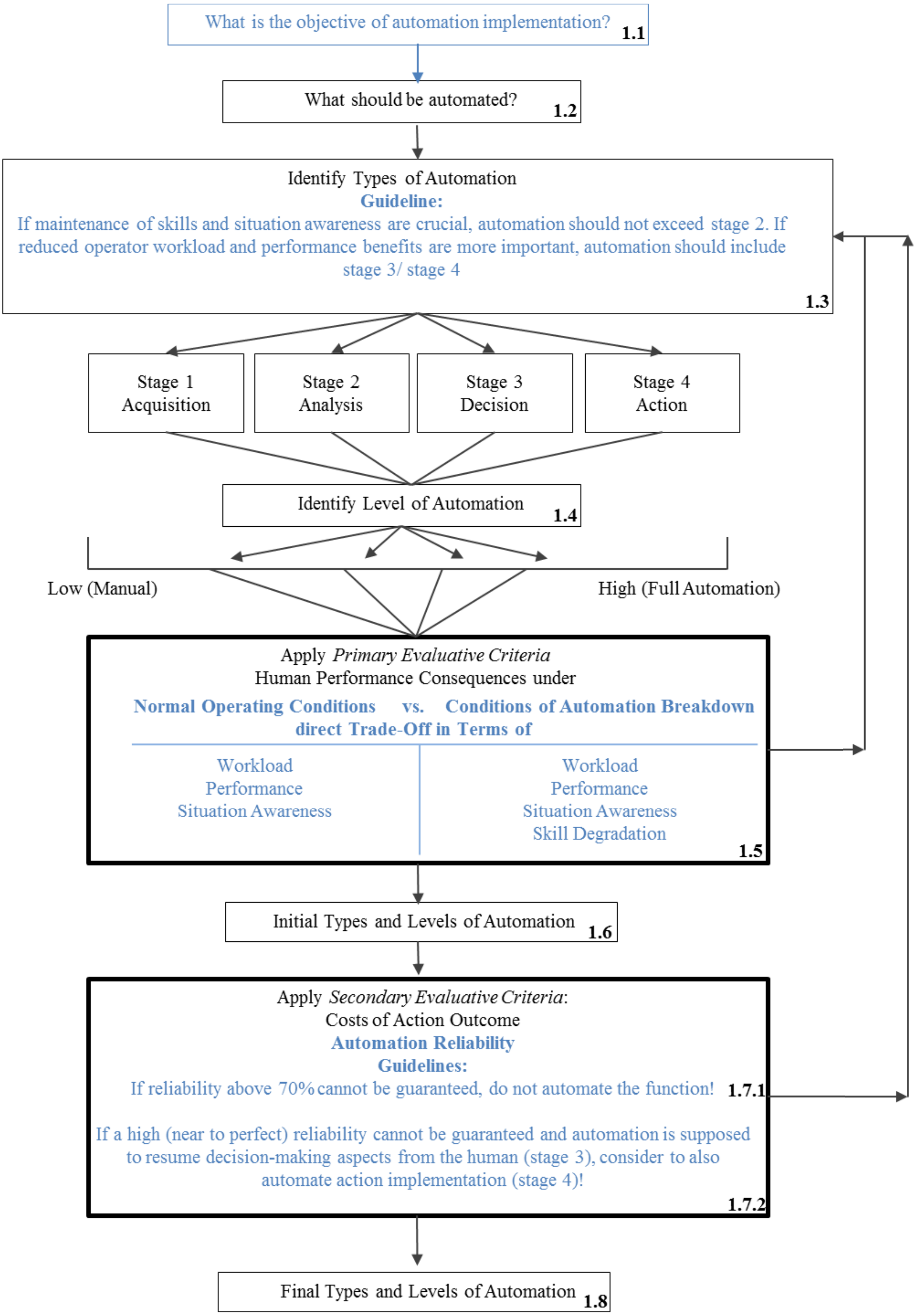

What is the objective of automation implementation?
1.1
What should be automated?
1.2
Identify Types of Automation
Guideline:
If maintenance of skills and situation awareness are crucial, automation should not exceed stage 2. If reduced operator workload and performance benefits are more important, automation should include stage 3/ stage 4
1.3
Stage 1
Acquisition
Stage 2
Analysis
Stage 3
Decision
Stage 4
Action
Identify Level of Automation
1.4
Low (Manual)
High (Full Automation)
Apply *Primary Evaluative Criteria*
Human Performance Consequences under
Normal Operating Conditions
vs.
Conditions of Automation Breakdown
direct Trade-Off in Terms of
Workload
Performance
Situation Awareness
Workload
Performance
Situation Awareness
Skill Degradation
1.5
Initial Types and Levels of Automation
1.6
Apply *Secondary Evaluative Criteria*:
Costs of Action Outcome
Automation Reliability
Guidelines:
If reliability above 70% cannot be guaranteed, do not automate the function!
1.7.1
If a high (near to perfect) reliability cannot be guaranteed and automation is supposed to resume decision-making aspects from the human (stage 3), consider to also automate action implementation (stage 4)!
1.7.2
Final Types and Levels of Automation
1.8

Figure 1. Flow chart model of automation design proposed by Parasuraman et al. (2000), extended based on results of this thesis. Specifications as well as extensions are highlighted in blue.

5.2.1 Specifications based on Study I – the Meta-Analysis

The meta-analysis revealed that an appropriate function allocation can only serve routine system performance and workload on the one hand, or maintenance of operators' manual skills and situation awareness on the other hand. This finding of a direct cost-benefit trade-off can be used to specify the first part of the flow chart model up to the primary evaluative criteria. As costs and benefits of automation trade-off directly, the consideration of the objective of automation implementation is crucial. Whereas under some circumstances an increase of productivity is aimed at, other main objectives may consider overall safety issues like a reduction of complexity for the operator. If the main objective is explicated, a weighting of possible costs and benefits of a certain function allocation, can be conducted that helps in deciding for an appropriate function allocation.

Therefore, the first question posed in the flow chart model should not ask what to automate, but why, i.e. with which objective:

What is the objective of automation implementation? (see figure 1, box 1.1)

With the objective of automation in mind, the question of what to automate can be posed (figure 1, box 1.2). Concerning the stage component of the levels and stages taxonomy, which is integrated in the flow chart model, results of the meta-analysis suggest the following: If operators' skill loss and the maintenance of situation awareness are of serious concern, for example, when an immediate manual take-over of automated functions under conditions of automation breakdown is relevant to safety, they should be kept involved at least to some extent in decision-making (stage 3), as well as action implementation (stage 4). Although, risks of skill degradation might not be fully excluded, they can probably be kept on a lower level. If however, skill degradation is not of major concern, for example, if fast manual interventions in case of automation breakdown might not be crucial, then automation exceeding stage 2 is preferable. Benefits like reduced workload and an enhanced overall performance under normal operating conditions are more likely compared with lower stage automation. Therefore, based on findings of the meta-analysis a first guideline referring to the stages of automation can be explicated. The guideline is integrated in the third step of the flow chart model, the identification of types of automation:

If maintenance of skills and situation awareness are crucial, automation should not exceed stage 2. If reduced operator workload and performance benefits are more important, automation should include stage 3/ stage 4. (see figure 1, box 1.3)

If the stages and levels of automation are initially defined, consequences of this combination have to be evaluated in detail. In the original model, the primary evaluative criteria are composed of a collection of concepts (mental workload, situation awareness, complacency, skill degradation) that are not further specified in relation to function allocation or system state. However, the consequences of automation should be explicitly differentiated for normal operating conditions and conditions under automation breakdown as this sub-division particularly contrasts costs and benefits (see figure 1, box 1.5).

Under normal operating conditions more automation leads to better performance, lower operator workload but also to reduced situation awareness. In terms of *automation breakdown* performance is negatively related to the degree of automation pointing to possible operators' skill degradations. Workload does not seem to be strongly affected by automation breakdown. However the latter finding needs further research as the power of this analysis was quite low (see chapter 2.4).

The conflicting effects of automation reveal that there is no specific configuration in which automation-induced performance benefits can be increased without any performance costs. Nevertheless, the specification of the flow chart model should support designers of automation to be aware of these trade-offs and decide for a more appropriate function allocation with regard to overall objectives and consequences for automation users.

When initial types and levels of automation have been identified and held out against the primary evaluation (figure 1, box 1.6), the secondary evaluative criteria have to be considered. However, Parasuraman et al. (2000) only mention aspects that have to be taken into account for this second evaluation instead of giving concrete advice for an appropriate function allocation. Therefore, findings of studies II and III supplement this part of the model.

5.2.2 Specifications based on Study II and Study III – the Laboratory Experiments

Concerning automation reliability, findings in both experimental studies show a clear boundary of reliability below which automation support is in fact not supportive. When reliability was realised below 70% participants behaved as if no automation support was available (study II and III). Attention allocation seemed to be completely unrelated to the automation's support whether this signalled a parameter deviation or not. In comparison to groups working with fairly reliabel automation (above 70%) these participants always spent most time on the automation-supported task (study II, differentiation of alarm and non-alarm periods). Therefore, automation less reliable than 70% did not reduce participants' cognitive effort to fulfil the task. Moreover, performance was also negatively affected by automation undershooting this boundary. The second study revealed negative effects of unreliability in concurrent task performance because participants reallocated attention to the automation-supported task. The third study revealed negative effects for performance in the automation-supported task. Although participants allocated as much attention to this task as participants without automation support, they could not reach superior performance with automation that still detected 68.75% of all deviations for them. Based on these findings which provide clear evidence for the reliability boundary first stated by Wickens and Dixon (2007), the second evaluative criterion of reliability can be supplemented by a concrete guideline:

If reliability above 70% cannot be guaranteed, do not automate the function!

(see figure 1, box 1.7.1)

Furthermore, the third study raised another aspect related to consequences of reliability in combination with different stages of automation. Function allocation in this study did not affect performance or attention allocation. However, prior research has provided evidence for a more severe effect of unreliability in interaction with higher-stage automation than with lower-stage automation (Crocoll & Coury, 1990; Rovira, McGarry, & Parasuraman, 2007; Sarter & Schroeder, 2001). As was already described in the discussion of this experiment in detail, these deviant findings could be due to the operationalisation of the higher-stage automation condition that differed from previous research. Whereas those studies realised higher-stage automation as decision-making automation (stage 3), the third experiment implemented automation that resumed decision-making but also the action implementation component of the task (stage 4), therefore representing an automation of the entire task. When a task is completely automated the operator's role is that of a supervisor, i.e. he has to monitor the proper functioning of the automation and is accountable for the overall task. In contrast to that, the role definition of an operator in interaction with stage 3 automation might not be that clear-cut. Whereas the information analysis, interpretation and decision-making components are resumed by automation the human operator is still responsible for the implementation of an action that was suggested by automation. Therefore, an operator who is still in charge of action implementation may define his role more or less limited to this subtask and interpret the automation's advices as a directive that has to be executed. When the automation is not perfectly reliable, this role understanding might lead to negative consequences as the human operator does not supervise the automation's proper functioning; incorrect advices by automation are not detected and accordingly end in inappropriate actions carried out by the human. Therefore, when an automation of the decision-making component is sought, the automation has to be sufficiently reliable. If this cannot be guaranteed, designers should redefine the operators' role to that of a supervisor by increasing the stage of automation. A corresponding guideline extends the secondary evaluative criteria of the flow chart model:

If a high (near to perfect) reliability cannot be guaranteed and automation is supposed to resume decision-making aspects from the human (stage 3), consider to also automate action implementation (stage 4)! (see figure 1, box 1.7.2)

5.2.3 Benefits of the Flow Chart Specifications for Automation Designers

Regarding the overall structure of the flow chart model, Parasuraman et al. (2000) propose the application of their model to decisions of function allocation between human and automation as an iterative process. This is illustrated by the recursive loops that are integrated in the flow chart (figure 1). The three studies of this thesis provide new insight that may help to minimise these feedback steps based on more precise guidelines that help with decisions in the first place. For example, the guideline referring to the identification of types of automation may help designers to make a more appropriate initial decision on what stages of automation should be implemented. By doing so, the design process gets more efficient as less feedback loops are needed. If a reconsideration of an initial automation is however advisable, e.g. because a sufficient reliability cannot be guaranteed, the guidelines that are directly integrated in the secondary evaluation process provide designers with concrete suggestions how function allocation should be changed (e.g. automating the complete task instead of only automating up to the decision-making component).

Furthermore, the introduction of the fundamental question of the objective of automation helps designers to understand how automation may be used when implemented. This involves designers of automation not only on the executive level, developing the automation, but provides them with a broader context and understanding of the desired outcome of automating certain functions (Parasuraman & Riley, 1997).

Moreover, the differentiation of primary evaluative criteria to normal-operating conditions and conditions of automation breakdown can be considered as an implicit guideline to evaluation studies of prototypes. The differentiation underlines the need to always consider and examine both conditions when applying those criteria.

In sum, the proposed extensions of the flow chart model not only improve the outcome of the design process, the function allocation between human and automation, but also the process itself by supporting an efficient application of the model.

5.3 Critical Considerations

The current thesis comprised two different methodical approaches. Regarding function allocation, the study's objective was to derive an overall picture of human performance consequences induced by automation based on results of available single studies to this topic. Therefore, a meta-analytic approach was chosen for study I. The objectives of studies II and III were to gain further insight into effects of different levels of reliability and combinations of reliability and function allocation. As prior research has rarely varied reliability over a whole range of levels and also rarely considered differential effects of function allocation and reliability at all, both studies were conducted as laboratory experiments. In the following, strengths and limitations of both methodical approaches shall be discussed separately.

5.3.1 Critical Considerations Concerning the Meta-Analysis

The objective of meta-analyses is to use data from different studies to obtain information about the effect size for a certain treatment on various constructs. In a classic meta-analytic approach estimates of effect sizes from the single studies are pooled to obtain an estimate of the averaged effect size across studies (Hedges, 1982). According to Rosenthal (1995) the two most common types of effect sizes are *r* and *d*. Examples for the *r* family are Pearson's product-moment correlations (*r*) or Fisher's *r*-to-*z* transformation. The most important *d* effect sizes are Cohen's *d*, Hedges' *g*, and Glass's Δ.

The meta-analysis conducted within the thesis had to deviate from the classic approach. This was done because effect sizes were only rarely reported in the original studies and other estimates of effect size like the F ratios for multiple conditions would not have been able to capture the correlational aspect of data. However, this was of particular relevance for the research question (the “lumber jack effect”). Therefore, an alternative coding of effects was applied which is discussed in the following.

The first challenge was to find a way to standardise the various forms of automation examined in the single studies. Considering that “more automation” can be achieved both by higher levels within a stage and/ or later stages (which, in the literature, are typically preceded by automation at earlier stages), it was assumed that increasing the two dimensions of Parasuraman et al.`s taxonomy (2000) increases the degree of automation (DOA). Accordingly, a clear-cut dominance rule was applied (for details see chapter 2.2) to convert combinations of stages and levels into rank data with an increasing rank reflecting an increasing DOA.

In a further step, effects of DOA on the meta-variables had to be classified as common effect sizes were not available. As an alternative solution, effects of DOA were coded dichotomous as significant or non-significant within the single studies. By doing so, the impact of DOA on those variables was expressed using rank orderings. Different rankings were assigned when there was a significant effect between two DOA conditions ($p < .05$). In case of non-significant effects between different degrees of automation tied ranks were assigned. A limitation of this relatively coarse-grained approach is that trends might have been underestimated. However, the data base did not allow for a more detailed coding.

The resulting ordinal scaling level made it necessary to use Kendall’s *tau* as an alternative analysis to Pearson’s product-moment correlations in the following analyses. With this non-parametric measure it was possible to determine and test for a monotonic relation between two dependent variables (e.g. DOA and workload). Furthermore, Kendall’s *tau* does not make the implicit assumption of equidistance between different rankings which would not have been the case for the data.

Although this is an unconventional approach, the chosen form of data aggregation is still in line with the basic idea of meta-analyses (Rosenthal, 1995). It is also in line with other authors who also departed from the classic approach for similar reasons (e.g. Hutchins, Wickens, Carolan, & Cumming, 2013; Wickens & Dixon, 2007; Wickens, Hooey, Gore, Sebok, & Koenicke, 2009; Wickens, Hutchins, Carolan, & Cumming, 2013).

However, a consequence that has to be kept in mind is that rank orders based on dominance orderings neither allowed for quantifying DOA on a ratio nor interval scale. Particularly, the observed trade-off, i.e. increasing automation leads to increased routine performance and reduced workload but also to an increased risk of skill degradation and loss of situation awareness, could not be related to specific levels or stages of automation in this first analysis step. As DOA was always defined relatively within a single study, the same ranking did not necessarily describe the same form of automation. For example, a higher DOA ranking within a study could be represented both by higher levels within a stage, or later stages. However, additionally conducting post-hoc analyses including studies that explicitly varied the stage component of automation yet allowed for a reference of results at least to this automation characteristic.

Certainly a superior approach would have assigned specific values to each type of automation thereby enabling a quantification of automation. Unfortunately, such a universal metric is not available yet, and it is doubted that it will be available in the near future given the countless instances of automation in miscellaneous contexts.

Besides the discussed limitations, a strength of the study is the definition of dependent variables. These were chosen with regard to the main aspects that are typically affected by automation implementation in terms of possible benefits and costs. Dependent variables were defined broad enough to group the data of single studies while still representing a clear definition of the concept in question. With this approach, various measures were merged into a single variable, which also implied an aggregation of different evaluation levels. Particularly, the meta-variables for workload and situation awareness not only included subjective measures typically quantified by the NASA-TLX (Hart & Staveland, 1988) or the SAGAT (Endsley, 1988; Endsley & Garland, 2000) but also included objective measures like secondary task performance for workload or any operators’ errors like mode errors (Sarter, 2008), omission or commission errors (Mosier & Skitka, 1996) for situation awareness. Thus, the generalisability and external validity of results increased.

In sum, the realised approach was considered as an appropriate analysis to provide a first access to an overall picture of impacts of function allocation. Particularly with regard to the data base which made a classic meta-analysis impractical, the overall pattern of raw effects and statistical results revealed a quantitative summary that has not been presented before in this detail and certainly offers an invitation for future research.

5.3.2 Critical Considerations Concerning the Laboratory Experiments

To evaluate impacts of reliability and function allocation on operator performance and cognitive demands the experimental task had to meet certain criteria. First, it had to be ensured that findings of the experiments can be generalised at least in main features to real world settings. Therefore, the experimental task should represent the complexity of real world settings while controlling for confounding factors. A typical working condition in complex human-machine systems is the requirement to work on several tasks concurrently. In this regard, laboratory experiments should provide a setting in which task fulfilment contains working on several tasks at the same time. Task fulfilment should be possible when working manually on all tasks; however, it should induce a high amount of cognitive effort. In this context, an implementation and use of automation in at least one of those tasks provides a coherent, realistic, and beneficial support.

Another important criterion that had to be met is the possibility to manipulate automation reliability and function allocation as independent variables within the experimental setting. Additionally the experimental setting should allow for an evaluation of dependent measures according to the single tasks that have to be performed. To evaluate cognitive demands in terms of attention allocation, eye-tracking measures had to be applicable to the experimental setting.

Last but not least, while fulfilling all the aforementioned criteria the experimental task had to be feasible for novices, i.e. for participants that have no prior experience with the task. Regarding economic aspects, this also implies that the experimental task further had to be easy to learn in a short period.

The Multi Attribute Task Battery (MATB; Comstock & Arnegard, 1992; Miller, 2010) meets all these criteria. The version which was used for the two experiments is a multi-task flight simulation, which consists of a two-dimensional compensatory tracking, engine-system monitoring, and fuel resource management that have to be performed concurrently. While the tasks of the MATB are easy to learn they still provide a challenging job that mirrors the cognitive demands of real world settings. Automation reliability could be manipulated to any level by implementing misses of the system in the monitoring task. Furthermore, function allocation could be varied. In this respect, automation could refer to only information acquisition in terms of alerting participants when a deviation in one of the parameters occurred or automation could resume the whole monitoring task by detecting malfunctions and resetting the affected parameter automatically.

A drawback of the MATB is the fact that it is not possible to determine a normative model for performance and attention allocation (Moray & Inagaki, 2000). The evaluation of dependent variables did not allow for conclusions of good or bad performance, nor did it allow interpretations if observed attention allocation strategies were appropriate or not. To overcome this deficiency, manual control groups were added to the experimental designs. These allowed evaluating dependent measures in relation to performance and attention allocation when no automation support was available. Therefore, (the extent of) benefits and costs of automation could be unambiguously determined.

A possible limitation of the conducted studies concerns the sample of participants. These were exclusively students from one of the three universities of Berlin, mainly from the TU Berlin. The acquisition of participants from a student population is a common approach in research domains like psychology or human factors. However, this fact might limit the representativeness of results to real world settings as differences in the population of students and real operators in terms of skills or attitudes cannot be excluded (Brewer & Crano, 2014).

Concerning study II, the generalisability of results may further be limited to a certain type of automation: information automation (stage 1 and 2, respectively). The automation implemented in this study represented a binary alarm system. Therefore, it could not be clarified to what extent results were transferable to more complex automation like decision automation. Correction of this flaw was objective of the third study. Results successfully supported findings of study II. However, only two very extreme forms of automation were compared: automation that supports the subtask of information acquisition (again a binary alarm system) and automation that resumes the whole task. This was done to ensure that they were discriminative enough to induce differences in terms of operators' performance and cognitive demands to be able to examine interaction effects in combination with reliability. However, as has already been discussed in detail in chapter 4.5, this realisation of automation support just might have accounted for non-findings on the impact of function allocation on performance and attention. Because the interaction with automation that resumes an entire task represents a special form of interaction, i.e. that of a supervisor with a subordinate, it might not be representative for all kinds of decision automation (particularly stage 3 automation). Therefore, results have to be interpreted with caution. Further research should investigate not only if reliability has differential effects for different stages of automation but more precisely should systematically vary automation stage by stage to gain detailed insight into effects on operators' performance and cognitive demands.

With reference to experimental design, both studies were conducted with a between-subjects design. For study II this resulted in five independent groups, four experimental conditions (different reliability levels) and one manual control group. Study III included four experimental groups (combinations of function allocation and reliability levels) and two manual control groups. Although a within-subjects design could have revealed even stronger evidence concerning participants' adaptation to different reliability levels, this design alternative was rejected. Research has shown that the experience of a certain automation reliability is likely to influence subsequent behaviour (e.g. Bahner, Hüper, & Manzey, 2008; Manzey et al., 2012; Parasuraman & Manzey, 2010). Therefore, participants could have been affected by foregoing adaptation processes in subsequent reliability trials. Furthermore, a within-subjects design could have the unintended effect that participants do not perceive different reliability levels of automation but rather average over the trials they have done so far. Therefore, a between-subjects design was seen as the appropriate methodical decision to avoid possible confounding influences and to assure that participants perceived the reliability of “their” automation appropriately. This was also assured by a manipulation check. In both studies findings confirmed that the manipulation of reliability had worked successfully as the perceived reliabilities were systematically related to the actual reliability levels and significantly differed between the experimental conditions. This was a precondition for all further analyses.

To evaluate the impact of reliability (and function allocation) on operator performance and cognitive demands dependent variables had to be defined. These were chosen in relation to the established hypotheses. Performance measures were evaluated separately for the three tasks participants had to work on. For the automation-supported monitoring task the dependent variable was defined as the joint performance of human and automation. This is in contrast to most studies that evaluated impacts of reliability on operator performance (e.g. Bailey & Scerbo, 2007; Parasuraman, Molloy, & Singh, 1993; Wiegmann, Rich, & Zhang, 2001). These studies evaluated performance in an automation-supported task solely by operator performance. This does not seem to be appropriate as the concept of an automated assistance is to support the operator and to resume parts of the task; i.e. the task is performed jointly. As a consequence it was considered important to respect the joint human-automation performance while evaluating overall performance benefits or costs associated with this sort of automated support.

The additional introduction of eye-tracking measures as dependent variables is a further strength of the current studies. This supplementary evaluation level, i.e. eye-tracking data for visual attention allocation, complies with Moray’s and Inagaki’s (2000; Moray, 2003) assertion to evaluate operators’ performance not only by fault detection but first and foremost by an analysis of their attention allocation strategies. These variables were important because they could explain findings on the performance level as they provided an insight into the strategies participants applied when working on the concurrent tasks.

In sum, the methodical approaches in all three studies were considered to be appropriate to examine the research questions. Moreover, the chosen approaches resolved some critical problems of prior research (e.g. provision of a more consistent result pattern by applying a meta-analytic approach in study I, applying eye-tracking measures in studies II and III) and therefore provided a sound basis for an interpretation of results. In this regard, the knowledge gained from the reported studies offers an invitation to future research, which is addressed in the last part of this work.

5.4 Outlook

This thesis provides a first step towards a specification of the flow chart model regarding decisions of function allocation in automation design (Parasuraman et al., 2000). Going beyond, findings also constitute a foundation for further research on human-automation interaction. This last chapter will relate the most important findings of this thesis to new research topics starting with aspects of function allocation.

The finding that benefits and costs of automation directly trade off provides a sound basis for future research. Particularly, research on adaptive automation should take the proposed boundary between information automation and decision automation into account. Adaptive automation is defined as the "… rational assignment of system functions to human and machine on a real-time basis for workload management and performance optimization." (Hancock et al., 2013, p. 9). Most studies available so far, operationalise modes of adaptive automation as either a complete automation of the task or the complete adoption of the task by the human operator (e.g. Bailey, Scerbo, Freeman, Mikulka, & Scott, 2006; Kaber, Wright, Prinzel, & Clamann, 2005; Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2003). This would reflect the extremes in every function allocation framework (e.g. Sheridan & Verplank, 1978) and does not accommodate the notion that automation is not an all-or-none decision. Findings of the meta-analysis suggest that a variation between information automation (stage 1 and 2) and decision automation (stage 3 and 4) could be of special interest regarding an optimal exploitation of automation benefits while reducing costs to an acceptable level.

Another major finding of the current thesis concerns a second boundary: the reliability cut-off around 70% below which automation cannot be considered as supportive for overall system performance and operator cognitive demands. In the studies revealing this finding, reliability was exclusively defined by automation misses. However, particularly in interaction with information automation like alarm systems, false alarms are of serious concern (Breznitz, 1984; Dixon et al., 2007; Meyer, Wiczorek, & Günzler, 2014). This kind of automation error was not regarded in the current thesis. Nevertheless, there is evidence that the proposed reliability boundary also seems to apply to unreliability in terms of false alarms (e.g. Dixon & Wickens, 2006; Dixon et al., 2004, 2007; Wickens et al., 2005; Wickens & Dixon, 2007). Therefore, the critical boundary seems to be a robust effect. Further research should examine why operators seem to perceive automation not useful when it is less reliable than 70%. From an objective viewpoint, an automation support that detects up to 70% of malfunctions still increases overall performance and has the potential to unload users' mental demands of the supported task. However, participants in the second and third study did not take advantage of this potential workload reduction. If the rationale underlying the expressed behaviour is known this might provide insight into what is missing for operators to be able to benefit from automation.

Moreover, future studies could focus on the relation between operators' adaptation to automation reliability above 70%. This could provide further information on operators' sensitivity to reliability changes. A high sensitivity could be assumed when the relation of behaviour/ attention and reliability follows a linear function, i.e. the higher automation's reliability the less attention is allocated to automation. An alternative relation could reveal a stepwise adaptation process, which would imply further reliability boundaries. These boundaries would mark a range of reliability levels to which operators behave indifferently. A more elaborated knowledge about operators' adaptation to automation's characteristics could help in designing reliability research. For example, if a stepwise adaptation was revealed, future studies could use this information to define distinct reliability levels that affect participants' behaviour (definition of high, medium, low reliability that is perceived as such). Moreover a stepwise adaptation process could be relevant for automation designers. Nowadays, one main challenge in interaction with alarm systems is concerned with the reduction of false alarms. If studies revealed that operators adapt stepwise, i.e. behave indifferent towards a range of reliabilities, then benchmarks could be defined that have to be undershot to prevent, for example, an ignorance of alarms. On the other hand, reliability ranges could be used to define the minimum reduction of false alarms that is needed to imply a change in operators' behaviour.

The third main aspect of this thesis focussed on possible interaction effects of function allocation and automation reliability (Study III). As neither main effects of function allocation nor differential effects were revealed, further research is needed. As numerous studies have revealed at least an impact of function allocation on operator performance (e.g. Cummings & Mitchell, 2007; Manzey et al., 2012; Onnasch, Wickens, et al., 2014; Rovira et al., 2007), possible interaction effects with reliability cannot be excluded based on non-findings of the last study. Future research should examine impacts of reliability by systematically varying automation stage by stage to gain detailed insight into effects on operators' performance and cognitive demands. In combination with prior research, findings of this thesis point to differential effects not only regarding information and decision automation, but also to differences between automation resuming functions up to decision-making and an automation of an entire task. The automation of cognitive task components might be even more critical for human operators than automating the complete task. Therefore, following studies should investigate if the boundary of function allocation is a real boundary or rather constitutes some kind of "function allocation valley" concerning stage 3 automation in which negative consequences for human operators are most likely.

In conclusion, this thesis contributed to a sound understanding of human-automation interaction. Findings of single studies were combined to reveal more valid and representative results, thus providing an overall picture of impacts of function allocation on performance and cognitive demands. Concerning reliability, existing research was broadened in two ways: First, impacts of a fine grained range of reliability levels were examined, and second, not only impacts on operators' performance but also on attentional strategies, which shed light on underlying processes of performance outcomes. Moreover, the lastly presented study combined both automation characteristics to look for differential effects that are not predictable when considering just one factor, either function allocation or automation reliability.

Beyond that, results of the single studies were integrated in the flow chart model for automation design (Parasuraman et al., 2000) to specify the evaluation processes when looking for an appropriate function allocation between human and automation. Therefore, this thesis provided not only new theoretical insight and an incitation for future research, but also detailed guidelines that may support practitioners in effective and efficient automation design.

5.5 References

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 688–699.

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 in Ergonomics Science, 8*(4), 321–348.

Bailey, N. R., Scerbo, M. W., Freeman, F. G., Mikulka, P. J., & Scott, L. A. (2006). Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human Factors, 48*(4), 693–709.

Brewer, M. B., & Crano, W. D. (2014). Research design and issues of validity. In H. T. Reis & C.M. Judd (Eds.) *Handbook of research methods in social and personality psychology* (2nd ed., 11 – 26). Cambridge University Press.

Breznitz, S. (1984). *Cry-wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Comstock, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research*. National Aeronautics and Space Administration, Langley Research Center Hampton, VA (Retrieved from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19920007912_1992007912.pdf?origin=publication_detail).

Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the 34th Annual Meeting of the Human Factors & Ergonomics Society* (1524–1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Cummings, M. L., & Mitchell, P. J. (2007). Operator scheduling strategies in supervisory control of multiple UAVs. *Aerospace Science and Technology, 11*(4), 339–348.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48*(3), 474–486.

Dixon, S. R., Wickens, C. D., & Chang, D. (2004). Unmanned aerial vehicle flight control: False alarms versus misses. *Proceedings of the 48th Annual Meeting of the Human Factors & Ergonomics Society* (152–156). Santa Monica, CA: Human Factors and Ergonomics Society.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors, 49*(4), 564–572.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the 32nd Annual Meeting of the Human Factors & Ergonomics Society* (97–101). Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R., & Garland, D. J. (2000). *Situation awareness analysis and measurement*. CRC Press.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*(2), 381–394.

Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-automation interaction research: Past, present, and future. *Ergonomics in Design: The Quarterly of Human Factors Applications, 21*(2), 9–14.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Peter A. Hancock and N. Meshkati (Eds.), *Advances in Psychology* (52, 139–183). North-Holland.

Hedges, L. V. (1982). *Statistical methodology in meta-analysis*. ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J (Retrieved from http://eric.ed.gov/?id=ED227133).

Hutchins, S. D., Wickens, C. D., Carolan, T. F., & Cumming, J. M. (2013). The influence of cognitive load on transfer with error prevention training methods: A meta-analysis. *Human Factors, 55*(4), 854–874.

Kaber, D. B., Wright, M. C., Prinzel, L. J., & Clamann, M. P. (2005). Adaptive automation of human-machine system information-processing functions. *Human Factors, 47*(4), 730–741.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making, 6*(1), 57–87.

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors, 56*(5), 840–849.

Milgram, P., Rastogi, A., & Grodski, J. J. (1995). Telerobotic control using augmented reality. *Proceedings of the 4th IEEE International Workshop on Robot and Human Communication*, 21–29.

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 behaviour*. Technical report no. AFRL-RH-WP-TR-2010-0133. Air Force Research Lab, Wright-Patterson, OH (Retrieved from http://dodreports.com/pdf/ada537547.pdf).

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics, 31*(3), 175–178.

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

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other. *Automation and Human Performance: Theory and Applications*, 201–220.

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators' adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies, 72*, 772-782.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*(3), 476–488.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology, 3*(1), 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.

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.

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. *Human Factors, 45*(4), 601–613.

Riley, V. (1989). A general model of mixed-initiative human-machine systems. *Proceedings of the 33rd Annual Meeting of the Human Factors & Ergonomics Society* (124–128). Santa Monica, CA: Human Factors and Ergonomics Society.

Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications*, (19–35). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*(2), 183–192.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors, 49*(1), 76–87.

Sarter, N. (2008). Investigating mode errors on automated flight decks: Illustrating the problem-driven, cumulative, and interdisciplinary nature of human factors research. *Human Factors, 50*(3), 506–510.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors, 43*(4), 573–583.

Sheridan, T. B. (2000). Function allocation: Algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies, 52*(2), 203–216.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. (Technical Report). Cambridge, MA: MIT Man-Machine Systems Laboratory.

Wickens, C. D., Dixon, S. R., Goh, J., Hammer, B. (2005). Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. *Proceedings of the 13th International Symposium on Aviation Psychology* (919–923). Columbus, OH: Association of Aviation Psychology.

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.

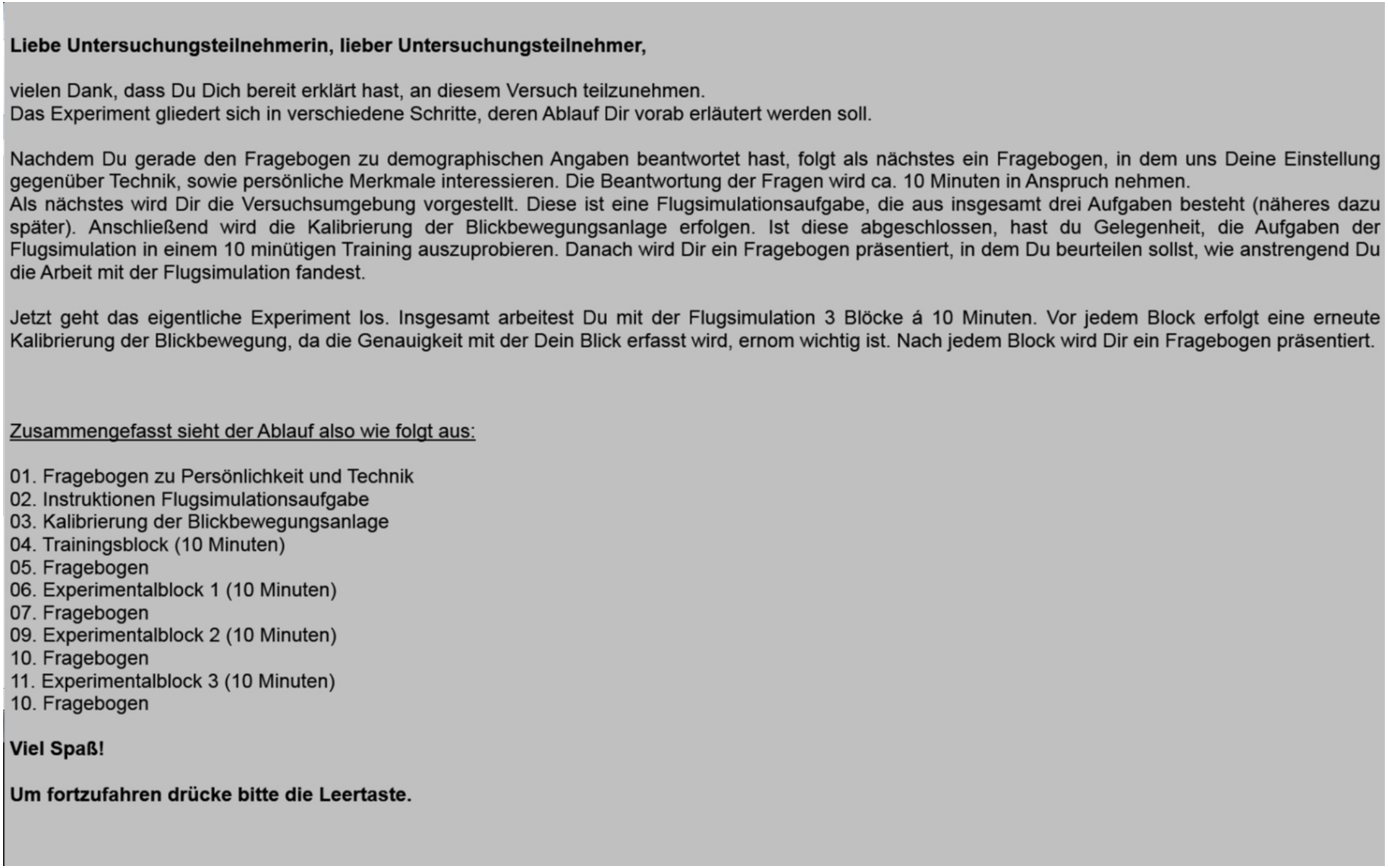
Wickens, C. D., Hooey, B. L., Gore, B. F., Sebok, A., & Koenicke, C. S. (2009). Identifying black swans in NextGen: Predicting human performance in off-nominal conditions. *Human Factors, 51*(5), 638–651.

Wickens, C. D., Hutchins, S., Carolan, T., & Cumming, J. (2013). Effectiveness of part-task training and increasing-difficulty training strategies: A meta-analysis approach. *Human Factors, 55*(2), 461–470.

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

Instructions Study II

Instructions were basically the same for the manual control group and the four automation-supported groups. When slides differed for a certain group, this is explicitly stated above the corresponding slide.



Liebe Untersuchungsteilnehmerin, lieber Untersuchungsteilnehmer,

vielen Dank, dass Du Dich bereit erklärt hast, an diesem Versuch teilzunehmen.
Das Experiment gliedert sich in verschiedene Schritte, deren Ablauf Dir vorab erläutert werden soll.

Nachdem Du gerade den Fragebogen zu demographischen Angaben beantwortet hast, folgt als nächstes ein Fragebogen, in dem uns Deine Einstellung gegenüber Technik, sowie persönliche Merkmale interessieren. Die Beantwortung der Fragen wird ca. 10 Minuten in Anspruch nehmen.
Als nächstes wird Dir die Versuchsumgebung vorgestellt. Diese ist eine Flugsimulationsaufgabe, die aus insgesamt drei Aufgaben besteht (näheres dazu später). Anschließend wird die Kalibrierung der Blickbewegungsanlage erfolgen. Ist diese abgeschlossen, hast du Gelegenheit, die Aufgaben der Flugsimulation in einem 10 minütigen Training auszuprobieren. Danach wird Dir ein Fragebogen präsentiert, in dem Du beurteilen sollst, wie anstrengend Du die Arbeit mit der Flugsimulation fandest.

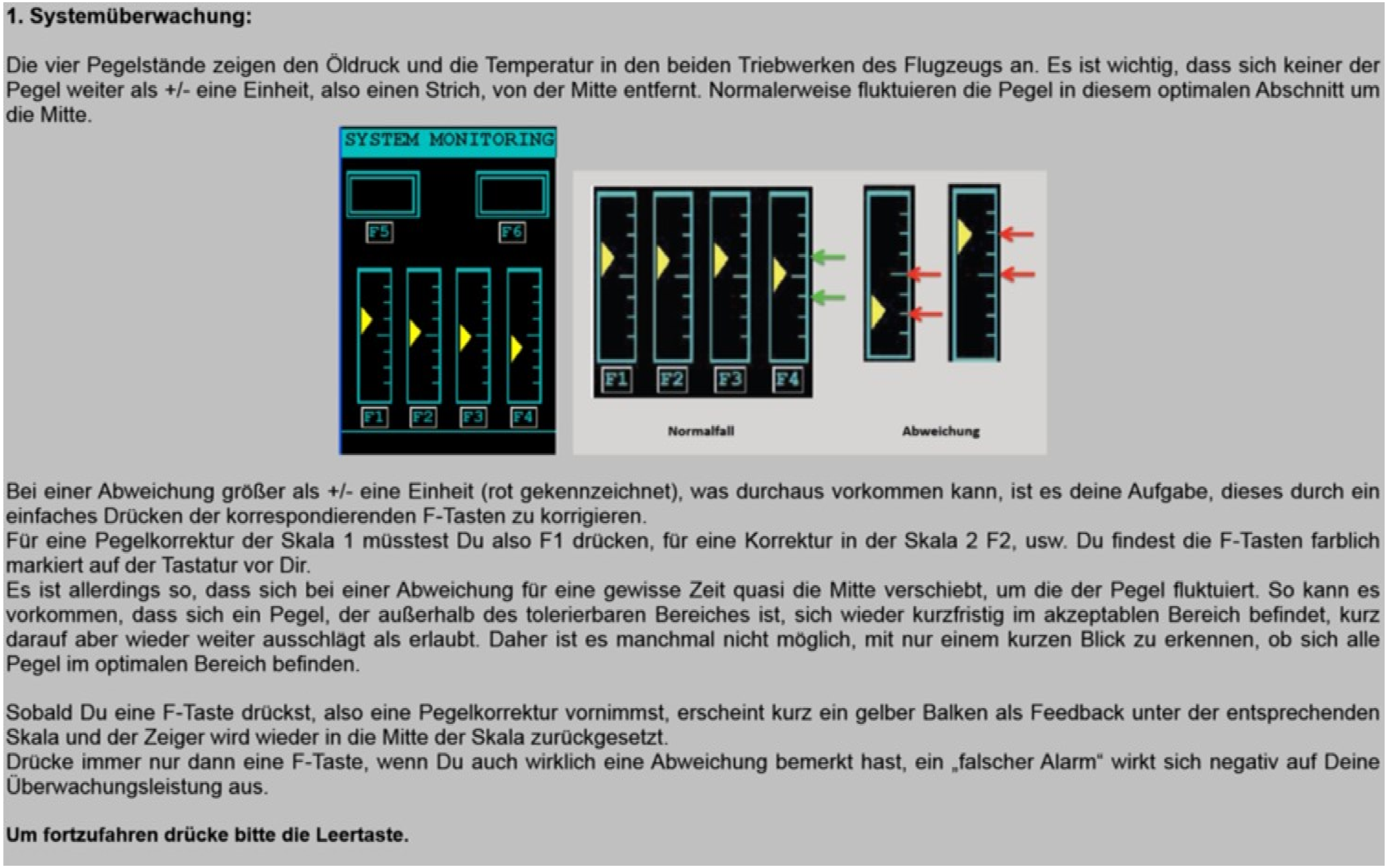
Jetzt geht das eigentliche Experiment los. Insgesamt arbeitest Du mit der Flugsimulation 3 Blöcke á 10 Minuten. Vor jedem Block erfolgt eine erneute Kalibrierung der Blickbewegung, da die Genauigkeit mit der Dein Blick erfasst wird, ernom wichtig ist. Nach jedem Block wird Dir ein Fragebogen präsentiert.

<u>Zusammengefasst sieht der Ablauf also wie folgt aus:</u>

01. Fragebogen zu Persönlichkeit und Technik
02. Instruktionen Flugsimulationsaufgabe
03. Kalibrierung der Blickbewegungsanlage
04. Trainingsblock (10 Minuten)
05. Fragebogen
06. Experimentalblock 1 (10 Minuten)
07. Fragebogen
09. Experimentalblock 2 (10 Minuten)
10. Fragebogen
11. Experimentalblock 3 (10 Minuten)
10. Fragebogen

Viel Spaß!

Um fortzufahren drücke bitte die Leertaste.



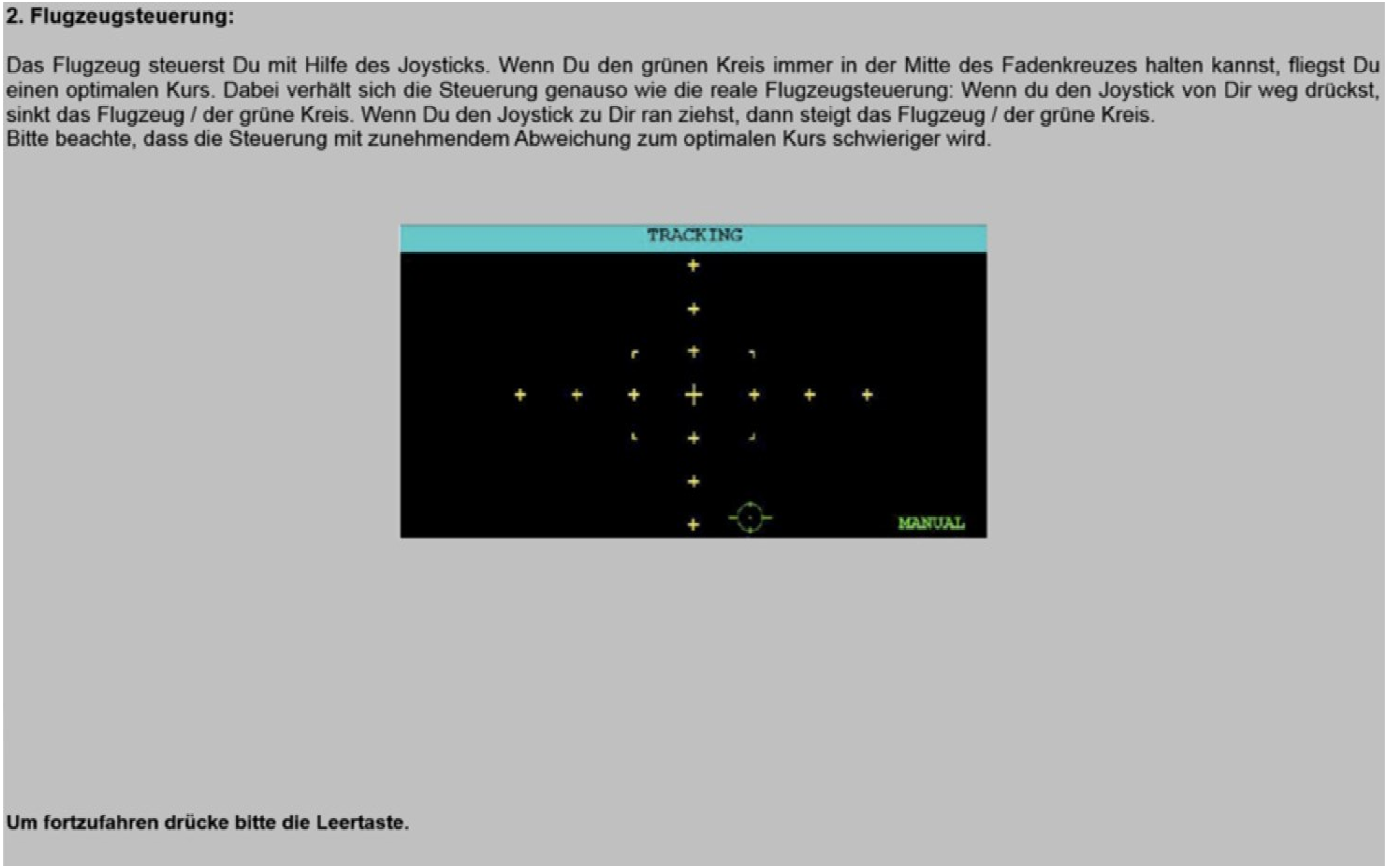
1. Systemüberwachung:

Die vier Pegelstände zeigen den Öldruck und die Temperatur in den beiden Triebwerken des Flugzeugs an. Es ist wichtig, dass sich keiner der Pegel weiter als +/- eine Einheit, also einen Strich, von der Mitte entfernt. Normalerweise fluktuieren die Pegel in diesem optimalen Abschnitt um die Mitte.

Bei einer Abweichung größer als +/- eine Einheit (rot gekennzeichnet), was durchaus vorkommen kann, ist es deine Aufgabe, dieses durch ein einfaches Drücken der korrespondierenden F-Tasten zu korrigieren.
Für eine Pegelkorrektur der Skala 1 müsstest Du also F1 drücken, für eine Korrektur in der Skala 2 F2, usw. Du findest die F-Tasten farblich markiert auf der Tastatur vor Dir.
Es ist allerdings so, dass sich bei einer Abweichung für eine gewisse Zeit quasi die Mitte verschiebt, um die der Pegel fluktuiert. So kann es vorkommen, dass sich ein Pegel, der außerhalb des tolerierbaren Bereiches ist, sich wieder kurzfristig im akzeptablen Bereich befindet, kurz darauf aber wieder weiter ausschlägt als erlaubt. Daher ist es manchmal nicht möglich, mit nur einem kurzen Blick zu erkennen, ob sich alle Pegel im optimalen Bereich befinden.

Sobald Du eine F-Taste drückst, also eine Pegelkorrektur vornimmst, erscheint kurz ein gelber Balken als Feedback unter der entsprechenden Skala und der Zeiger wird wieder in die Mitte der Skala zurückgesetzt.
Drücke immer nur dann eine F-Taste, wenn Du auch wirklich eine Abweichung bemerkt hast, ein „falscher Alarm“ wirkt sich negativ auf Deine Überwachungsleistung aus.

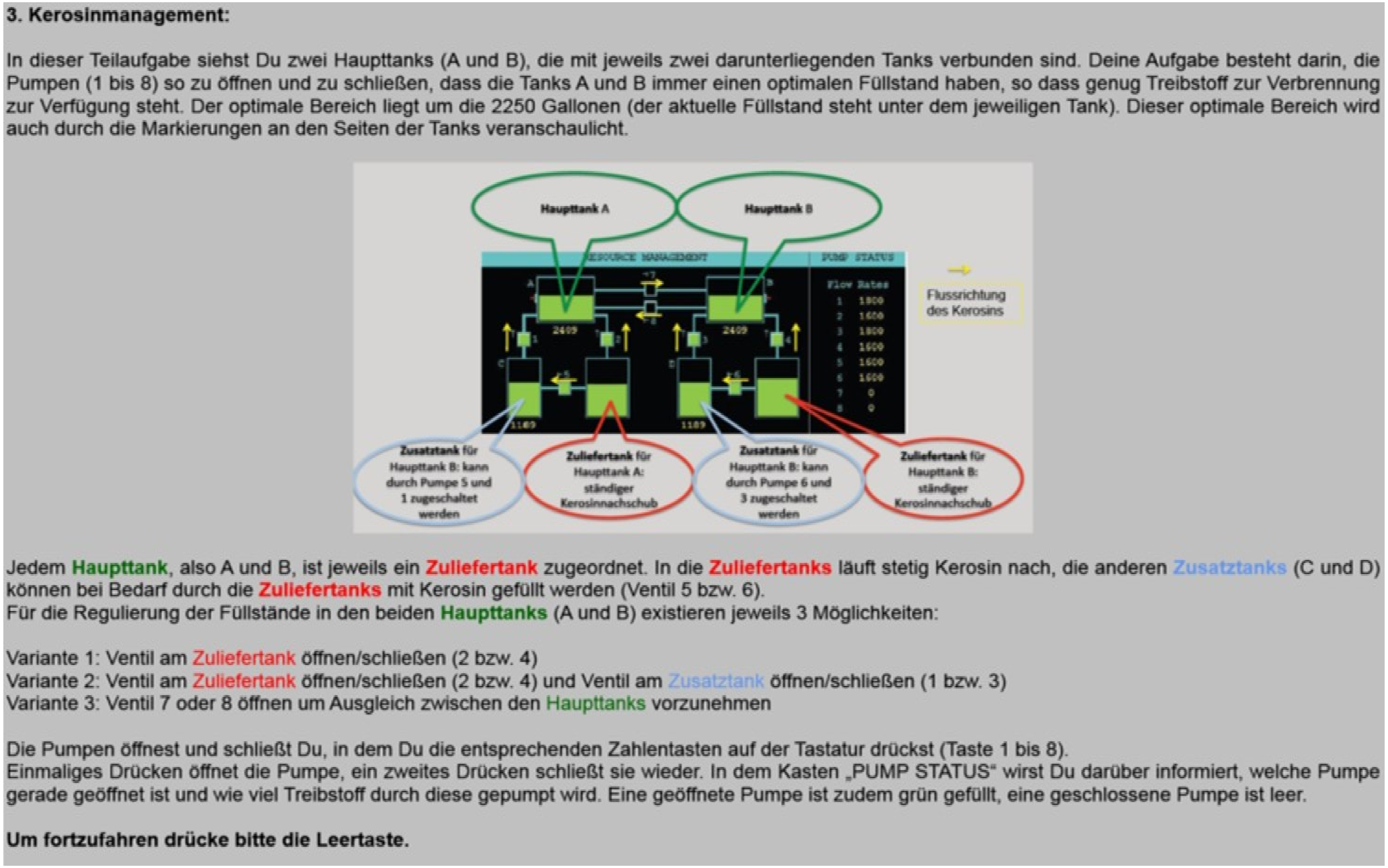
Um fortzufahren drücke bitte die Leertaste.



2. Flugzeugsteuerung:

Das Flugzeug steuerst Du mit Hilfe des Joysticks. Wenn Du den grünen Kreis immer in der Mitte des Fadenkreuzes halten kannst, fliegst Du einen optimalen Kurs. Dabei verhält sich die Steuerung genauso wie die reale Flugzeugsteuerung: Wenn du den Joystick von Dir weg drückst, sinkt das Flugzeug / der grüne Kreis. Wenn Du den Joystick zu Dir ran ziehst, dann steigt das Flugzeug / der grüne Kreis.
Bitte beachte, dass die Steuerung mit zunehmendem Abweichung zum optimalen Kurs schwieriger wird.

Um fortzufahren drücke bitte die Leertaste.



3. Kerosinmanagement:

In dieser Teilaufgabe siehst Du zwei Haupttanks (A und B), die mit jeweils zwei darunterliegenden Tanks verbunden sind. Deine Aufgabe besteht darin, die Pumpen (1 bis 8) so zu öffnen und zu schließen, dass die Tanks A und B immer einen optimalen Füllstand haben, so dass genug Treibstoff zur Verbrennung zur Verfügung steht. Der optimale Bereich liegt um die 2250 Gallonen (der aktuelle Füllstand steht unter dem jeweiligen Tank). Dieser optimale Bereich wird auch durch die Markierungen an den Seiten der Tanks veranschaulicht.

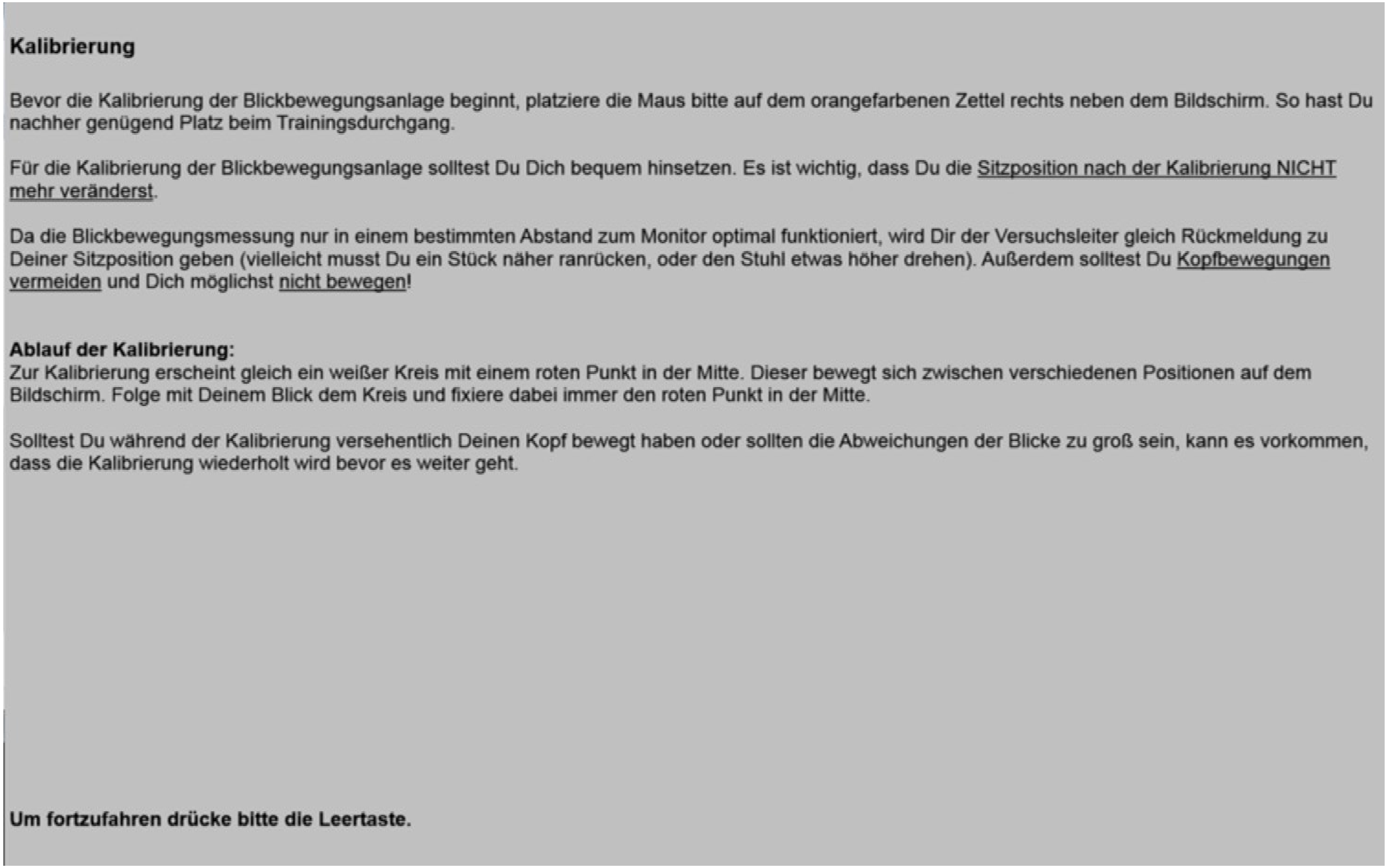
Haupttank A
Haupttank B
RESOURCE MANAGEMENT
PUMP STATUS
A
B
C
D
2409
2409
1109
1109
1
2
3
4
5
6
7
8
Flow Rates
1 1800
2 1600
3 1800
4 1600
5 1600
6 1600
7 0
8 0
Flussrichtung des Kerosins
Zusatztank für Haupttank B: kann durch Pumpe 5 und 1 zugeschaltet werden
Zuliefertank für Haupttank A: ständiger Kerosinnachschub
Zusatztank für Haupttank B: kann durch Pumpe 6 und 3 zugeschaltet werden
Zuliefertank für Haupttank B: ständiger Kerosinnachschub

Jedem **Haupttank**, also A und B, ist jeweils ein **Zuliefertank** zugeordnet. In die **Zuliefertanks** läuft stetig Kerosin nach, die anderen **Zusatztanks** (C und D) können bei Bedarf durch die **Zuliefertanks** mit Kerosin gefüllt werden (Ventil 5 bzw. 6).
Für die Regulierung der Füllstände in den beiden **Haupttanks** (A und B) existieren jeweils 3 Möglichkeiten:

Variante 1: Ventil am Zuliefertank öffnen/schließen (2 bzw. 4)
Variante 2: Ventil am Zuliefertank öffnen/schließen (2 bzw. 4) und Ventil am Zusatztank öffnen/schließen (1 bzw. 3)
Variante 3: Ventil 7 oder 8 öffnen um Ausgleich zwischen den Haupttanks vorzunehmen

Die Pumpen öffnest und schließt Du, in dem Du die entsprechenden Zahlentasten auf der Tastatur drückst (Taste 1 bis 8).
Einmaliges Drücken öffnet die Pumpe, ein zweites Drücken schließt sie wieder. In dem Kasten „PUMP STATUS“ wirst Du darüber informiert, welche Pumpe gerade geöffnet ist und wie viel Treibstoff durch diese gepumpt wird. Eine geöffnete Pumpe ist zudem grün gefüllt, eine geschlossene Pumpe ist leer.

Um fortzufahren drücke bitte die Leertaste.



Kalibrierung

Bevor die Kalibrierung der Blickbewegungsanlage beginnt, platziere die Maus bitte auf dem orangefarbenen Zettel rechts neben dem Bildschirm. So hast Du nachher genügend Platz beim Trainingsdurchgang.

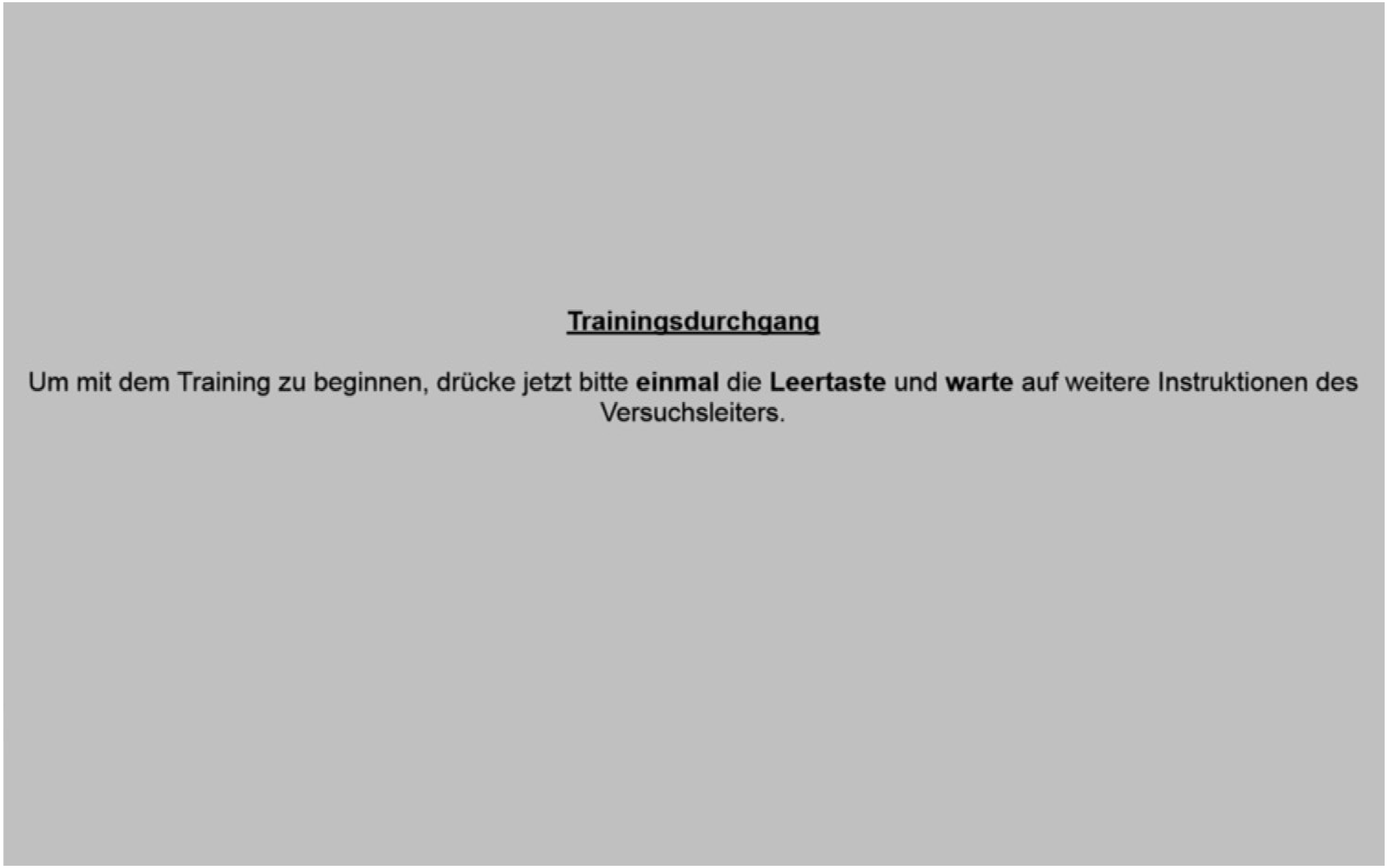
Für die Kalibrierung der Blickbewegungsanlage solltest Du Dich bequem hinsetzen. Es ist wichtig, dass Du die Sitzposition nach der Kalibrierung NICHT mehr veränderst.

Da die Blickbewegungsmessung nur in einem bestimmten Abstand zum Monitor optimal funktioniert, wird Dir der Versuchsleiter gleich Rückmeldung zu Deiner Sitzposition geben (vielleicht musst Du ein Stück näher ranrücken, oder den Stuhl etwas höher drehen). Außerdem solltest Du Kopfbewegungen vermeiden und Dich möglichst nicht bewegen!

Ablauf der Kalibrierung:
Zur Kalibrierung erscheint gleich ein weißer Kreis mit einem roten Punkt in der Mitte. Dieser bewegt sich zwischen verschiedenen Positionen auf dem Bildschirm. Folge mit Deinem Blick dem Kreis und fixiere dabei immer den roten Punkt in der Mitte.

Solltest Du während der Kalibrierung versehentlich Deinen Kopf bewegt haben oder sollten die Abweichungen der Blicke zu groß sein, kann es vorkommen, dass die Kalibrierung wiederholt wird bevor es weiter geht.

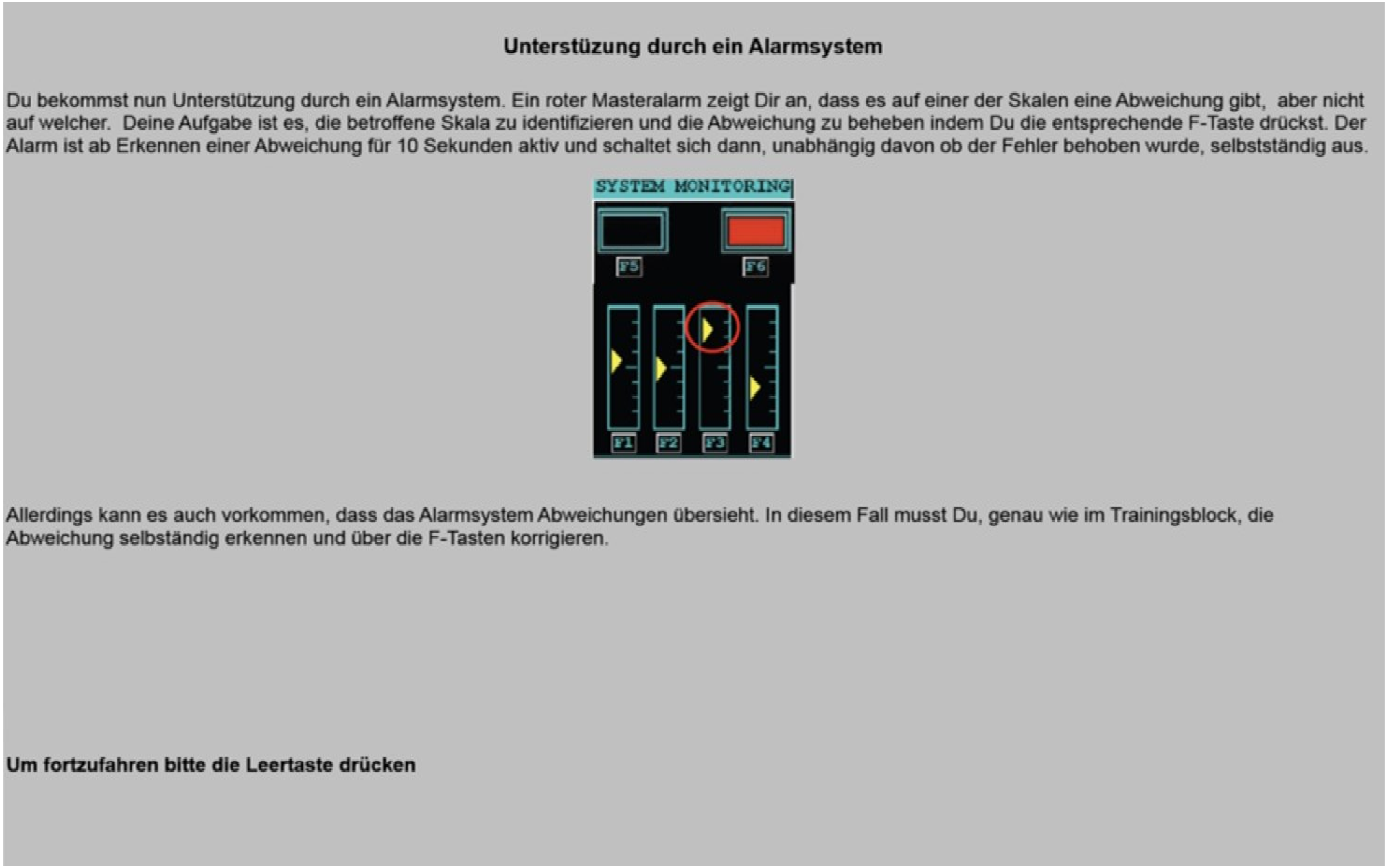
Um fortzufahren drücke bitte die Leertaste.



Trainingsdurchgang

Um mit dem Training zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

This slide was only provided to the automation-supported group.



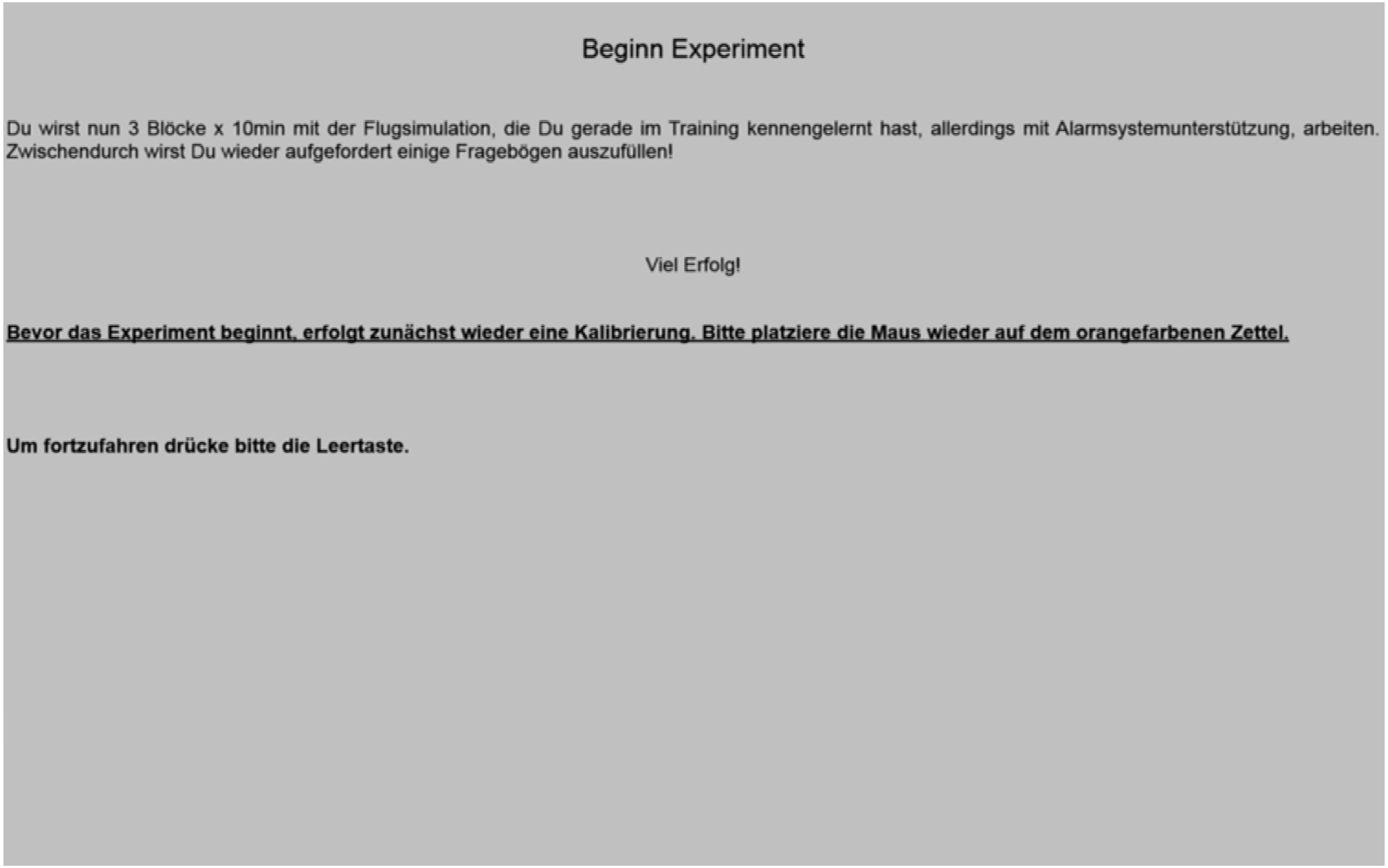
Unterstüzung durch ein Alarmsystem

Du bekommst nun Unterstützung durch ein Alarmsystem. Ein roter Masteralarm zeigt Dir an, dass es auf einer der Skalen eine Abweichung gibt, aber nicht auf welcher. Deine Aufgabe ist es, die betroffene Skala zu identifizieren und die Abweichung zu beheben indem Du die entsprechende F-Taste drückst. Der Alarm ist ab Erkennen einer Abweichung für 10 Sekunden aktiv und schaltet sich dann, unabhängig davon ob der Fehler behoben wurde, selbstständig aus.

Allerdings kann es auch vorkommen, dass das Alarmsystem Abweichungen übersieht. In diesem Fall musst Du, genau wie im Trainingsblock, die Abweichung selbständig erkennen und über die F-Tasten korrigieren.

Um fortzufahren bitte die Leertaste drücken

This slide was only provided to the automation-supported group.



Beginn Experiment

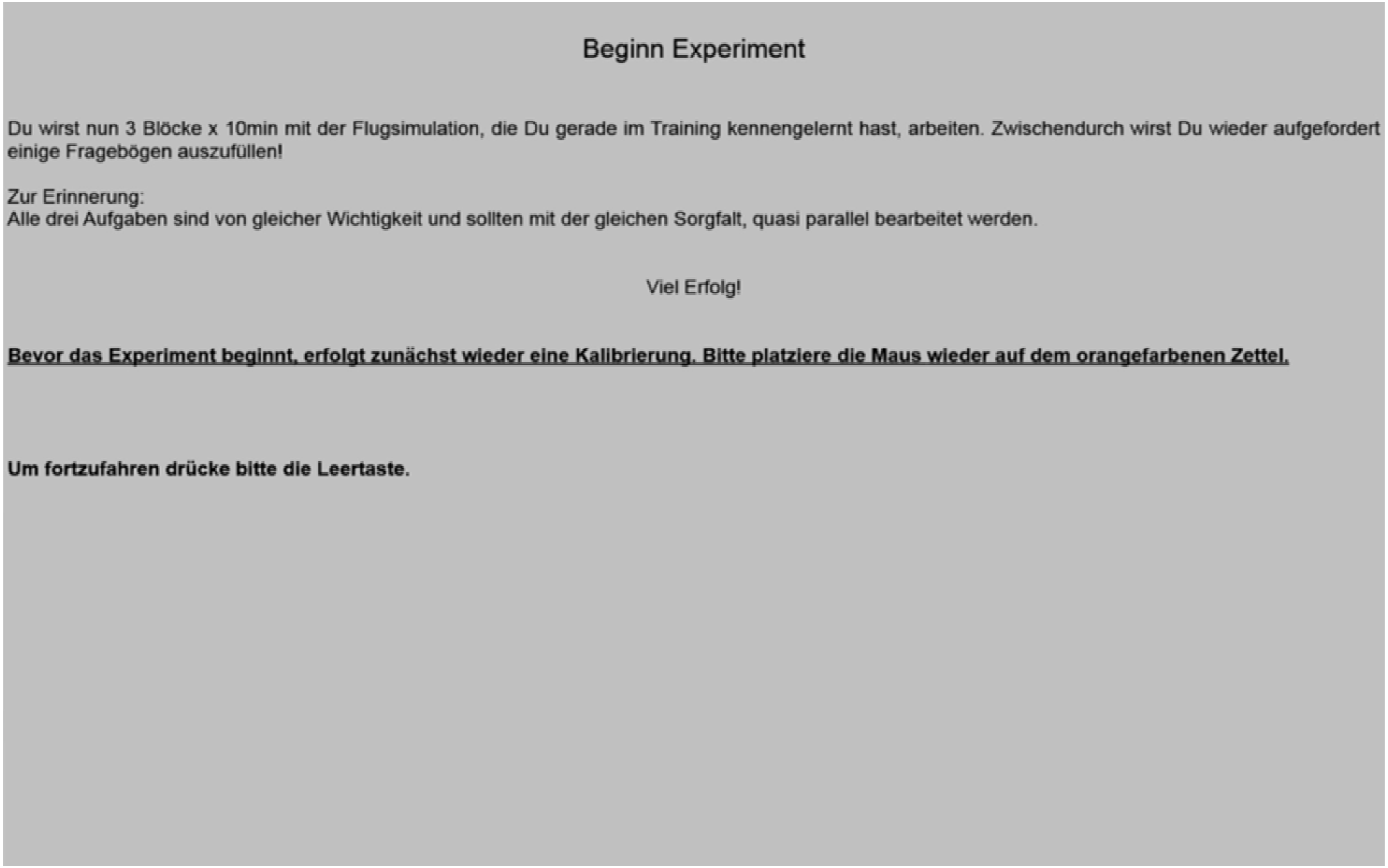
Du wirst nun 3 Blöcke x 10min mit der Flugsimulation, die Du gerade im Training kennengelernt hast, allerdings mit Alarmsystemunterstützung, arbeiten. Zwischendurch wirst Du wieder aufgefordert einige Fragebögen auszufüllen!

Viel Erfolg!

Bevor das Experiment beginnt, erfolgt zunächst wieder eine Kalibrierung. Bitte platziere die Maus wieder auf dem orangefarbenen Zettel.

Um fortzufahren drücke bitte die Leertaste.

This slide presents the equivalent for the manual control group.



Beginn Experiment

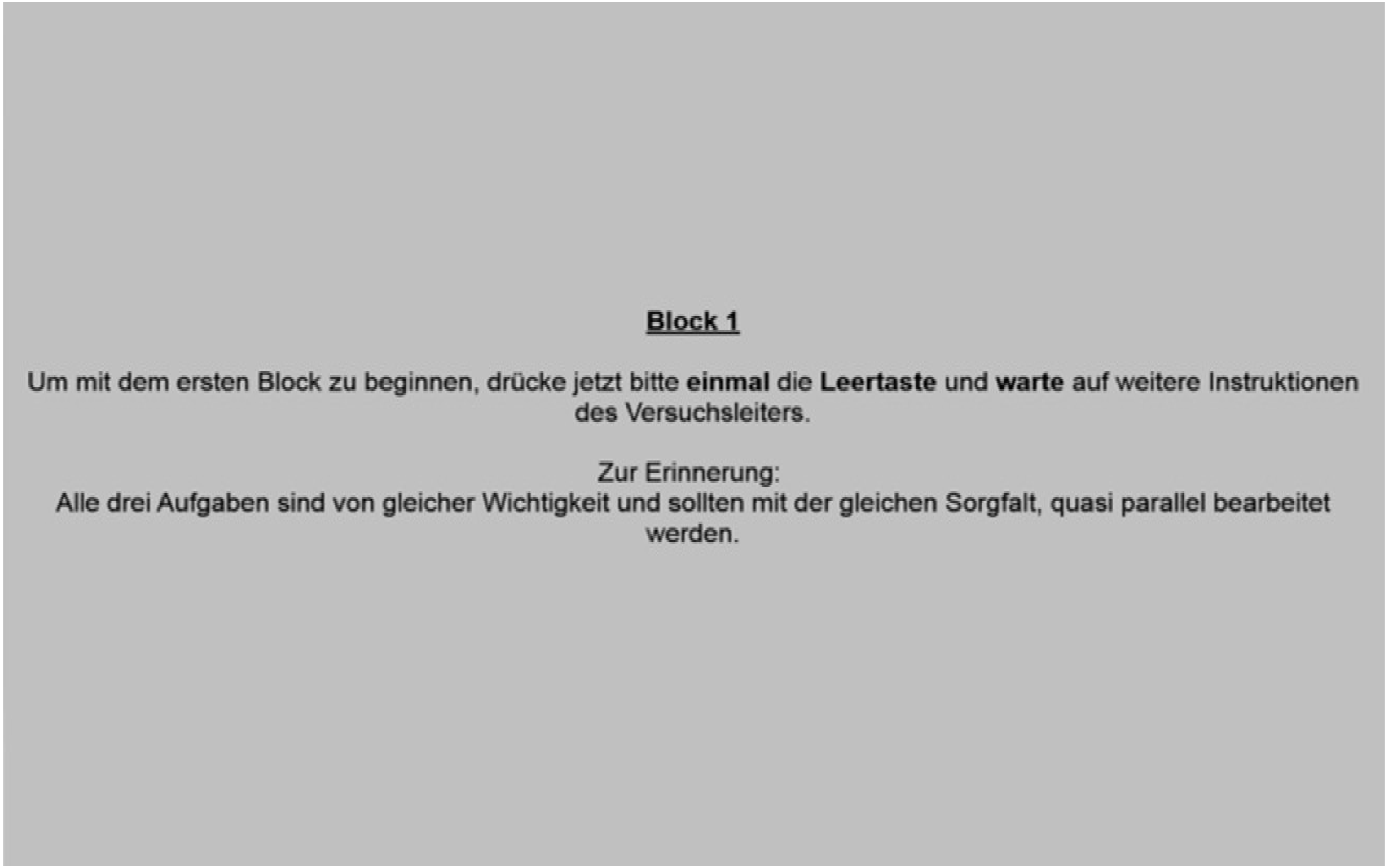
Du wirst nun 3 Blöcke x 10min mit der Flugsimulation, die Du gerade im Training kennengelernt hast, arbeiten. Zwischendurch wirst Du wieder aufgefordert einige Fragebögen auszufüllen!

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.

Viel Erfolg!

Bevor das Experiment beginnt, erfolgt zunächst wieder eine Kalibrierung. Bitte platziere die Maus wieder auf dem orangefarbenen Zettel.

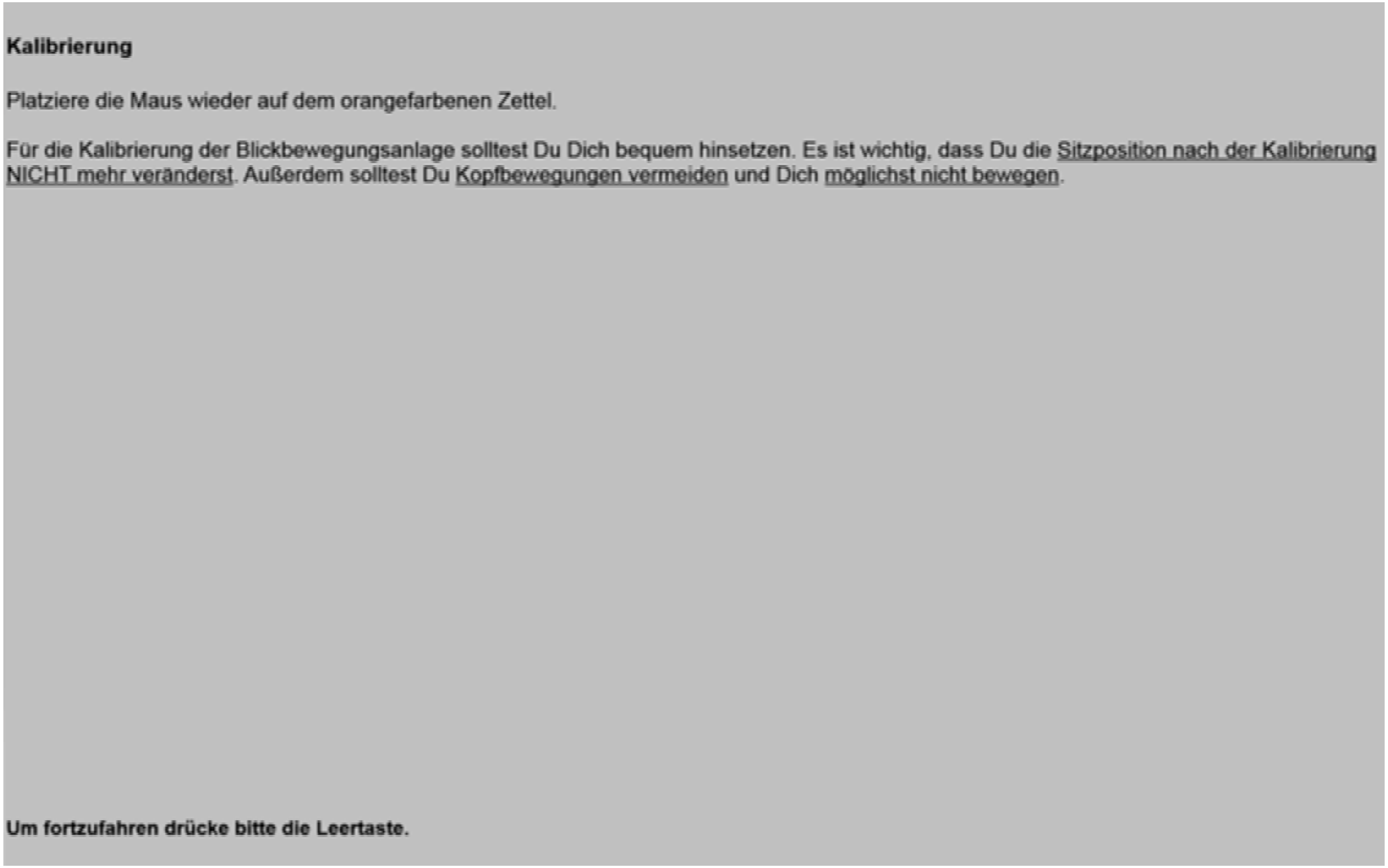
Um fortzufahren drücke bitte die Leertaste.



<u>Block 1</u>

Um mit dem ersten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.

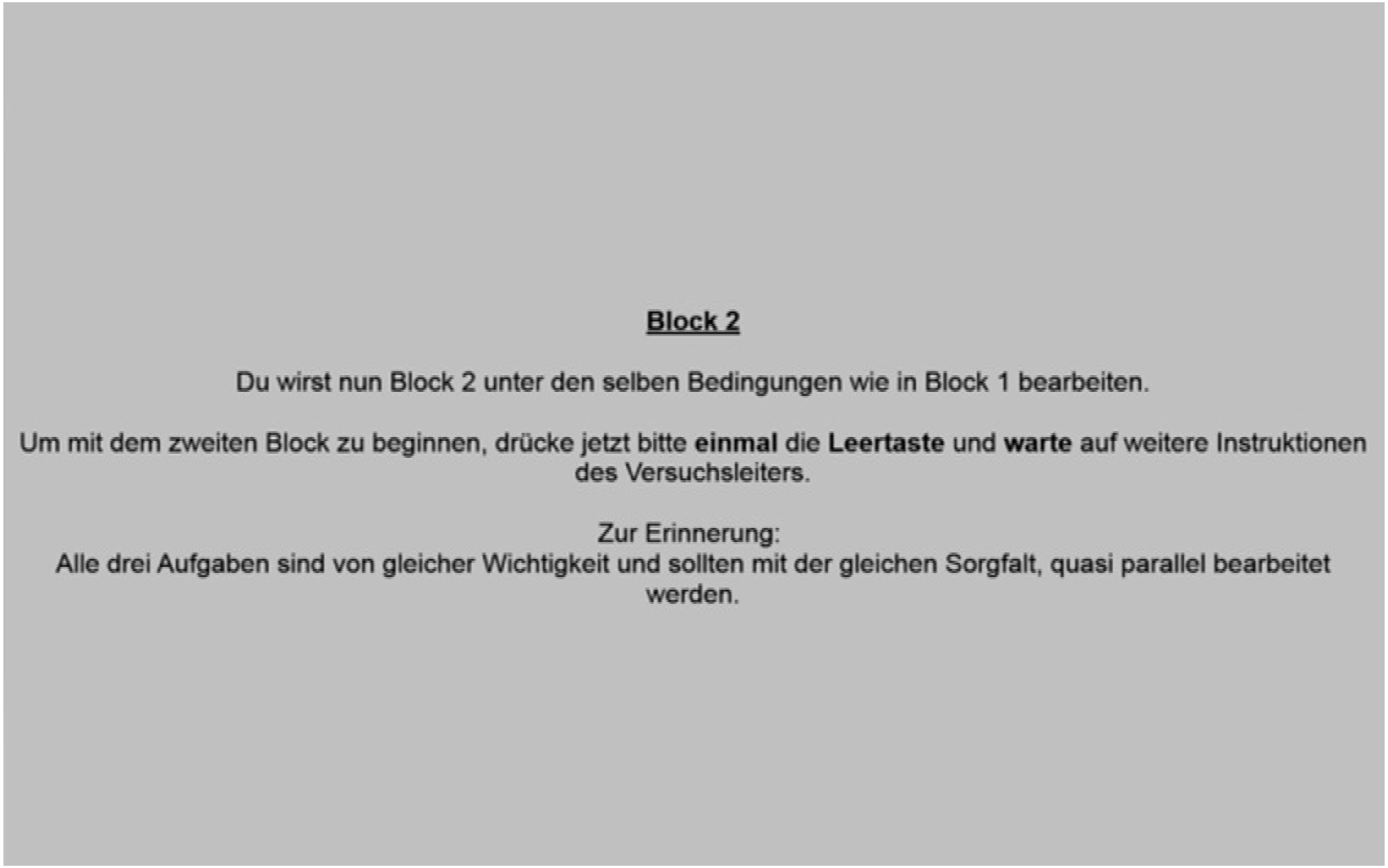


Kalibrierung

Platziere die Maus wieder auf dem orangefarbenen Zettel.

Für die Kalibrierung der Blickbewegungsanlage solltest Du Dich bequem hinsetzen. Es ist wichtig, dass Du die Sitzposition nach der Kalibrierung NICHT mehr veränderst. Außerdem solltest Du Kopfbewegungen vermeiden und Dich möglichst nicht bewegen.

Um fortzufahren drücke bitte die Leertaste.

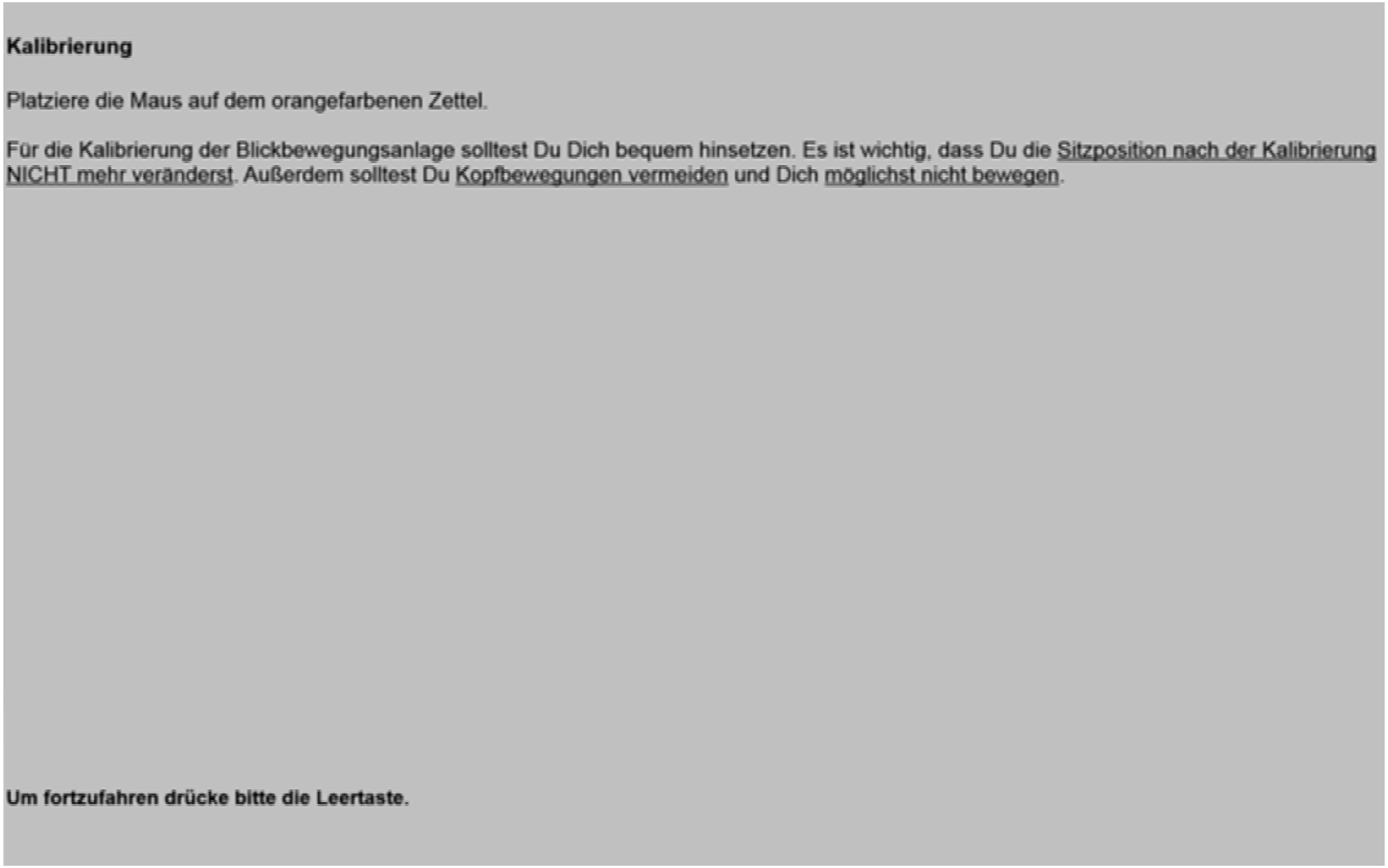


Block 2

Du wirst nun Block 2 unter den selben Bedingungen wie in Block 1 bearbeiten.

Um mit dem zweiten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.

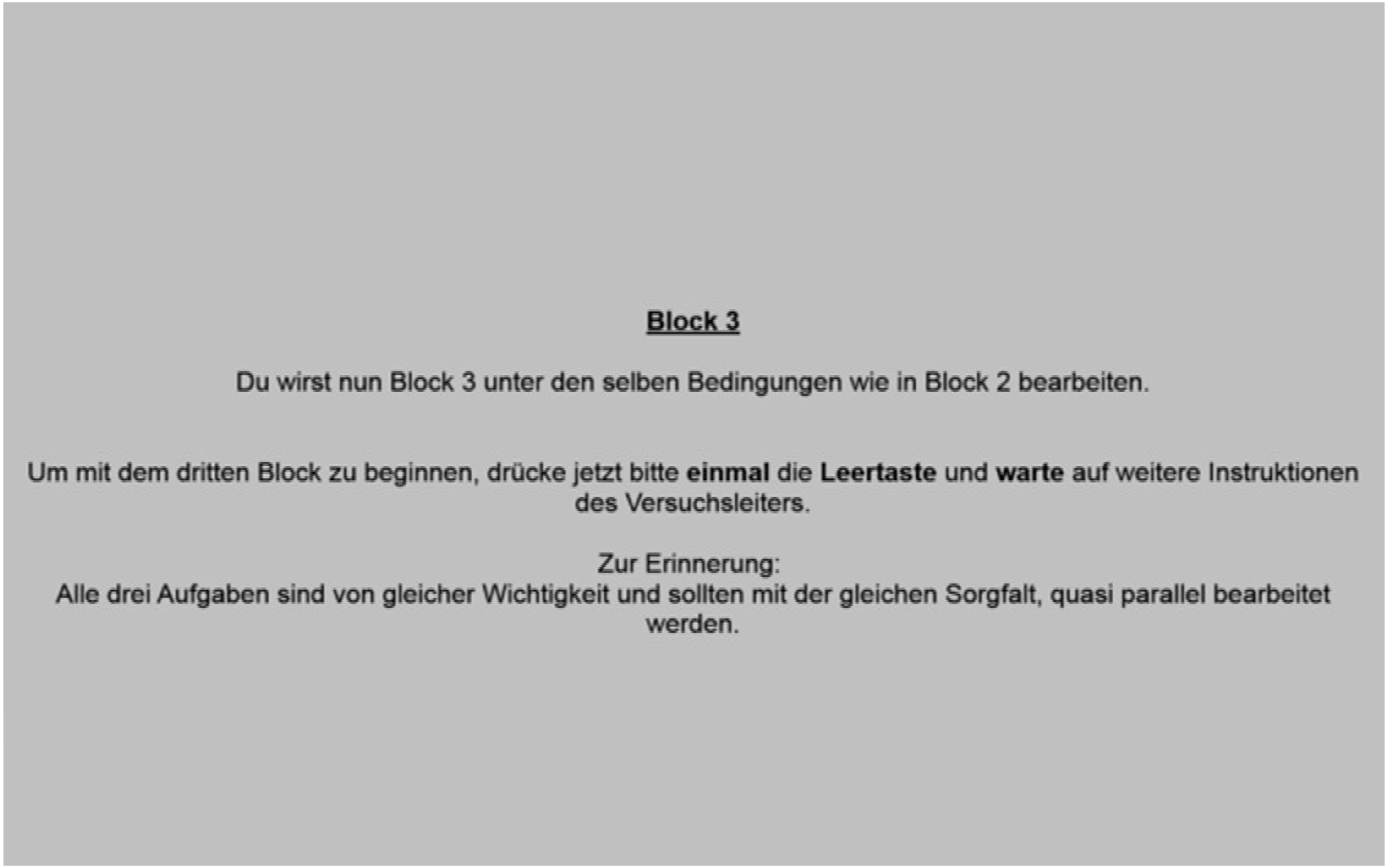


Kalibrierung

Platziere die Maus auf dem orangefarbenen Zettel.

Für die Kalibrierung der Blickbewegungsanlage solltest Du Dich bequem hinsetzen. Es ist wichtig, dass Du die Sitzposition nach der Kalibrierung NICHT mehr veränderst. Außerdem solltest Du Kopfbewegungen vermeiden und Dich möglichst nicht bewegen.

Um fortzufahren drücke bitte die Leertaste.

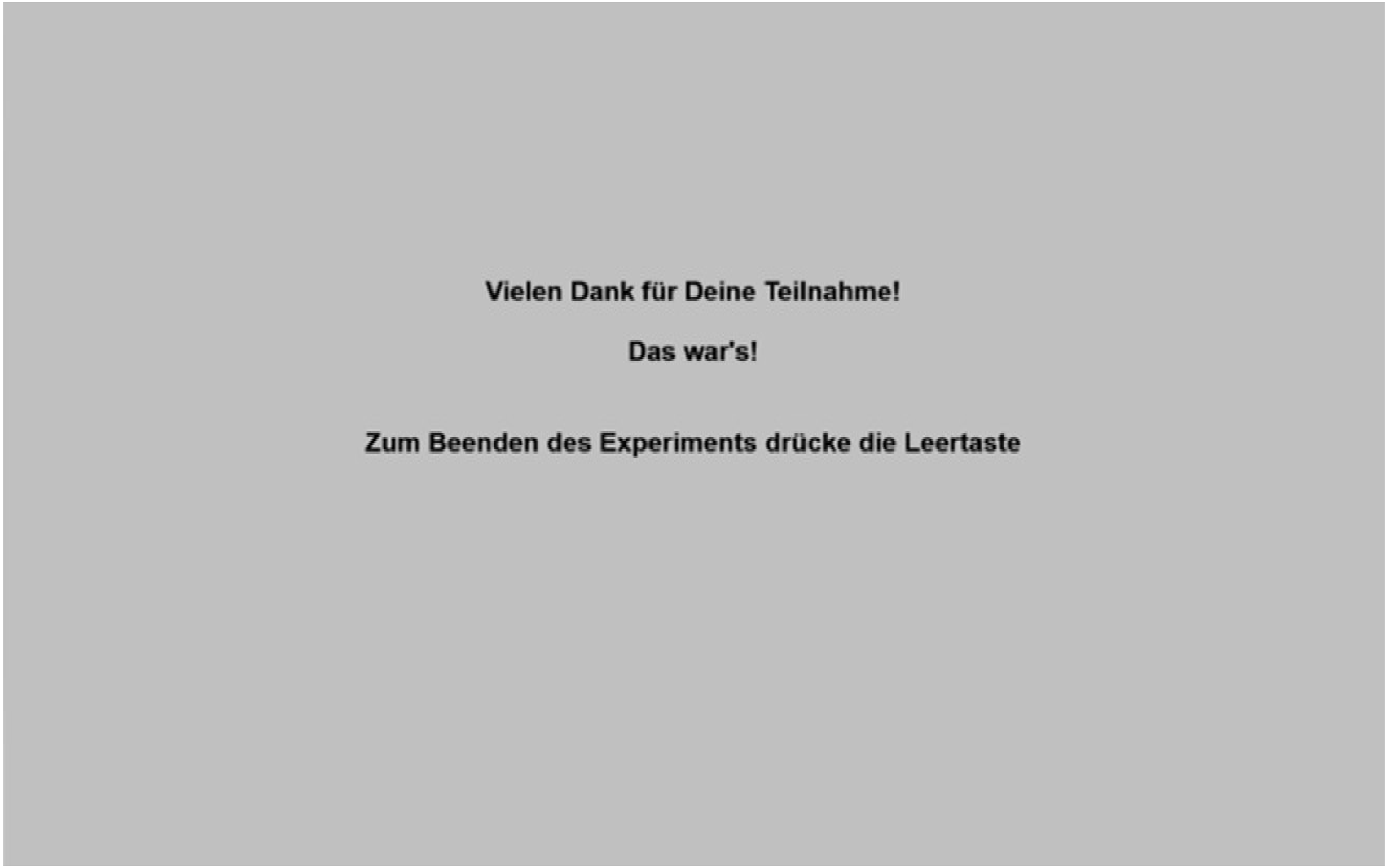


Block 3

Du wirst nun Block 3 unter den selben Bedingungen wie in Block 2 bearbeiten.

Um mit dem dritten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.



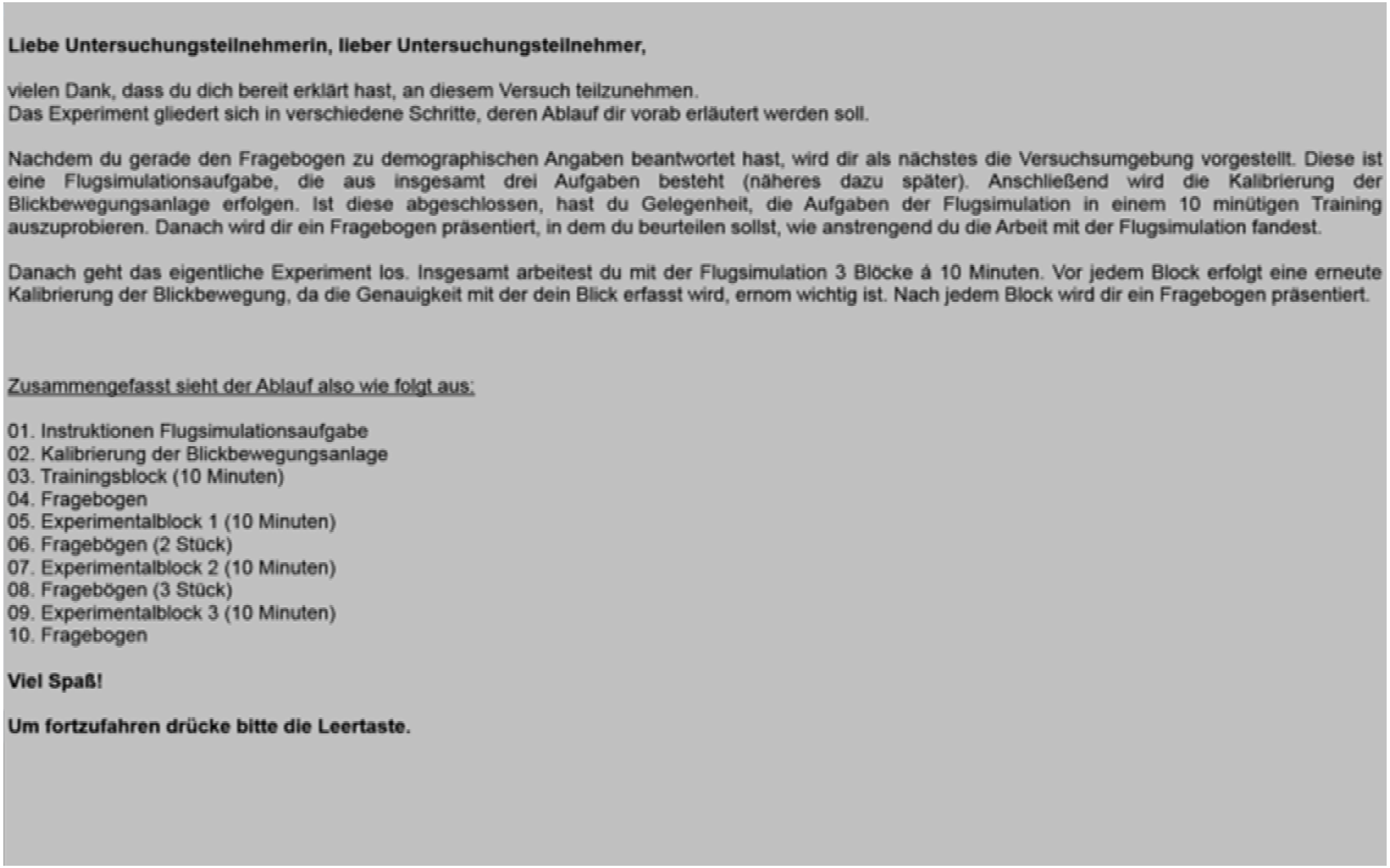
Vielen Dank für Deine Teilnahme!

Das war's!

Zum Beenden des Experiments drücke die Leertaste

Instructions Study III

Instructions were basically the same for the two manual control groups and the four automation-supported groups. When slides differed for a certain group, this is explicitly stated above the corresponding slide.



Liebe Untersuchungsteilnehmerin, lieber Untersuchungsteilnehmer,

vielen Dank, dass du dich bereit erklärt hast, an diesem Versuch teilzunehmen.
Das Experiment gliedert sich in verschiedene Schritte, deren Ablauf dir vorab erläutert werden soll.

Nachdem du gerade den Fragebogen zu demographischen Angaben beantwortet hast, wird dir als nächstes die Versuchsumgebung vorgestellt. Diese ist eine Flugsimulationsaufgabe, die aus insgesamt drei Aufgaben besteht (näheres dazu später). Anschließend wird die Kalibrierung der Blickbewegungsanlage erfolgen. Ist diese abgeschlossen, hast du Gelegenheit, die Aufgaben der Flugsimulation in einem 10 minütigen Training auszuprobieren. Danach wird dir ein Fragebogen präsentiert, in dem du beurteilen sollst, wie anstrengend du die Arbeit mit der Flugsimulation fandest.

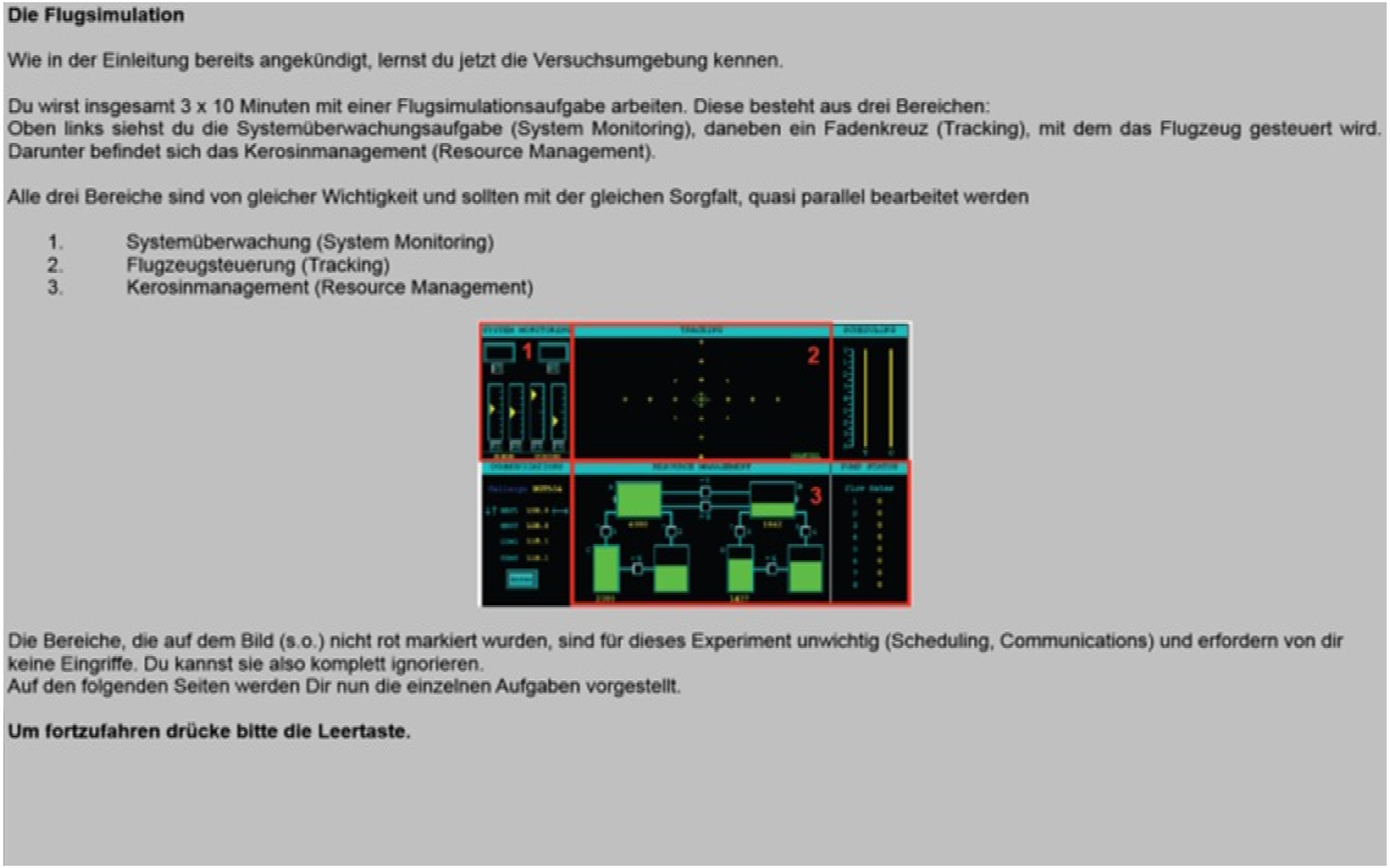
Danach geht das eigentliche Experiment los. Insgesamt arbeitest du mit der Flugsimulation 3 Blöcke á 10 Minuten. Vor jedem Block erfolgt eine erneute Kalibrierung der Blickbewegung, da die Genauigkeit mit der dein Blick erfasst wird, ernom wichtig ist. Nach jedem Block wird dir ein Fragebogen präsentiert.

<u>Zusammengefasst sieht der Ablauf also wie folgt aus:</u>

01. Instruktionen Flugsimulationsaufgabe
02. Kalibrierung der Blickbewegungsanlage
03. Trainingsblock (10 Minuten)
04. Fragebogen
05. Experimentalblock 1 (10 Minuten)
06. Fragebögen (2 Stück)
07. Experimentalblock 2 (10 Minuten)
08. Fragebögen (3 Stück)
09. Experimentalblock 3 (10 Minuten)
10. Fragebogen

Viel Spaß!

Um fortzufahren drücke bitte die Leertaste.



Die Flugsimulation

Wie in der Einleitung bereits angekündigt, lernst du jetzt die Versuchsumgebung kennen.

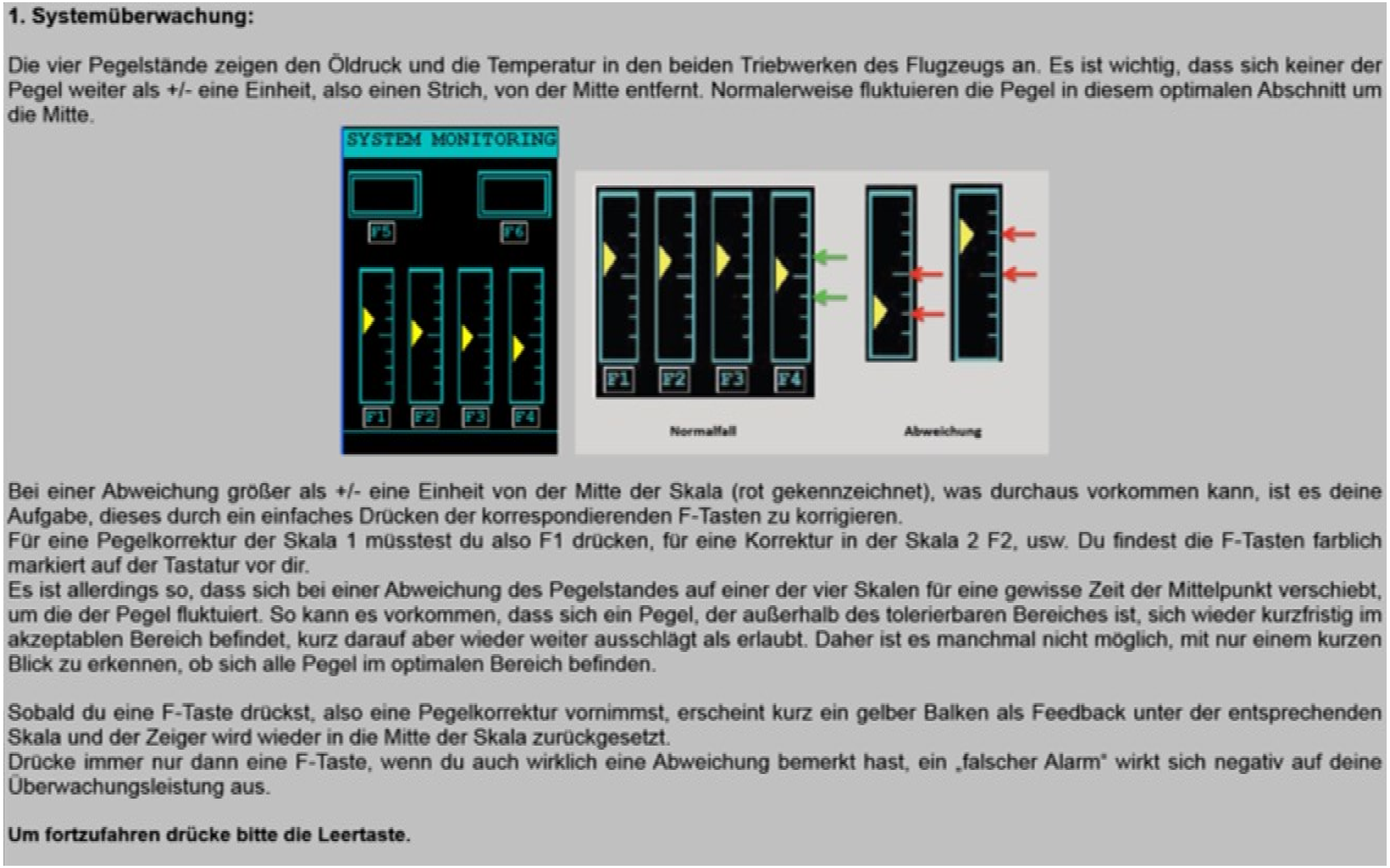
Du wirst insgesamt 3 x 10 Minuten mit einer Flugsimulationsaufgabe arbeiten. Diese besteht aus drei Bereichen:
Oben links siehst du die Systemüberwachungsaufgabe (System Monitoring), daneben ein Fadenkreuz (Tracking), mit dem das Flugzeug gesteuert wird. Darunter befindet sich das Kerosinmanagement (Resource Management).

Alle drei Bereiche sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden

1. Systemüberwachung (System Monitoring)
2. Flugzeugsteuerung (Tracking)
3. Kerosinmanagement (Resource Management)

Die Bereiche, die auf dem Bild (s.o.) nicht rot markiert wurden, sind für dieses Experiment unwichtig (Scheduling, Communications) und erfordern von dir keine Eingriffe. Du kannst sie also komplett ignorieren.
Auf den folgenden Seiten werden Dir nun die einzelnen Aufgaben vorgestellt.

Um fortzufahren drücke bitte die Leertaste.



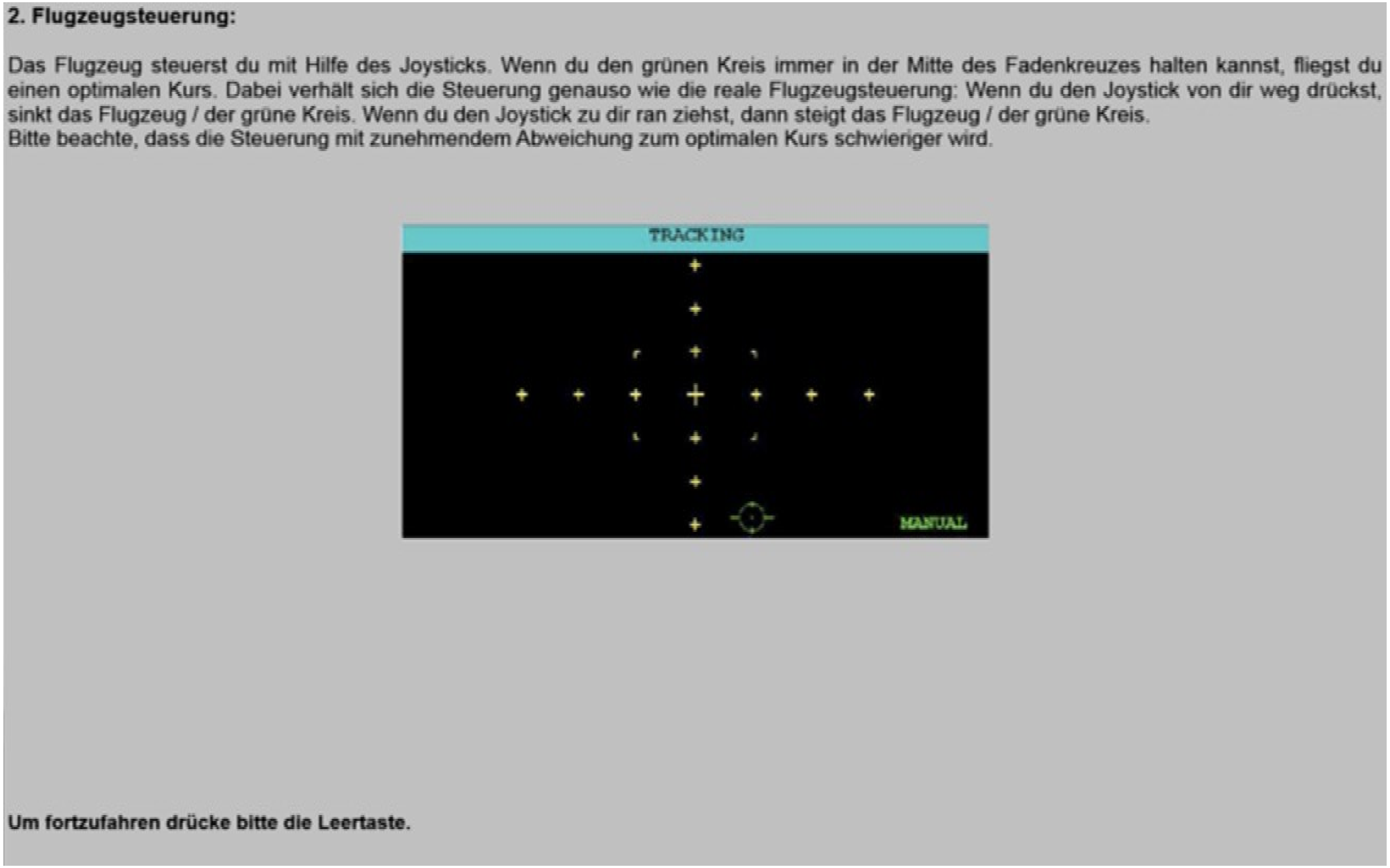
1. Systemüberwachung:

Die vier Pegelstände zeigen den Öldruck und die Temperatur in den beiden Triebwerken des Flugzeugs an. Es ist wichtig, dass sich keiner der Pegel weiter als +/- eine Einheit, also einen Strich, von der Mitte entfernt. Normalerweise fluktuieren die Pegel in diesem optimalen Abschnitt um die Mitte.

Bei einer Abweichung größer als +/- eine Einheit von der Mitte der Skala (rot gekennzeichnet), was durchaus vorkommen kann, ist es deine Aufgabe, dieses durch ein einfaches Drücken der korrespondierenden F-Tasten zu korrigieren.
Für eine Pegelkorrektur der Skala 1 müsstest du also F1 drücken, für eine Korrektur in der Skala 2 F2, usw. Du findest die F-Tasten farblich markiert auf der Tastatur vor dir.
Es ist allerdings so, dass sich bei einer Abweichung des Pegelstandes auf einer der vier Skalen für eine gewisse Zeit der Mittelpunkt verschiebt, um die der Pegel fluktuiert. So kann es vorkommen, dass sich ein Pegel, der außerhalb des tolerierbaren Bereiches ist, sich wieder kurzfristig im akzeptablen Bereich befindet, kurz darauf aber wieder weiter ausschlägt als erlaubt. Daher ist es manchmal nicht möglich, mit nur einem kurzen Blick zu erkennen, ob sich alle Pegel im optimalen Bereich befinden.

Sobald du eine F-Taste drückst, also eine Pegelkorrektur vornimmst, erscheint kurz ein gelber Balken als Feedback unter der entsprechenden Skala und der Zeiger wird wieder in die Mitte der Skala zurückgesetzt.
Drücke immer nur dann eine F-Taste, wenn du auch wirklich eine Abweichung bemerkt hast, ein „falscher Alarm" wirkt sich negativ auf deine Überwachungsleistung aus.

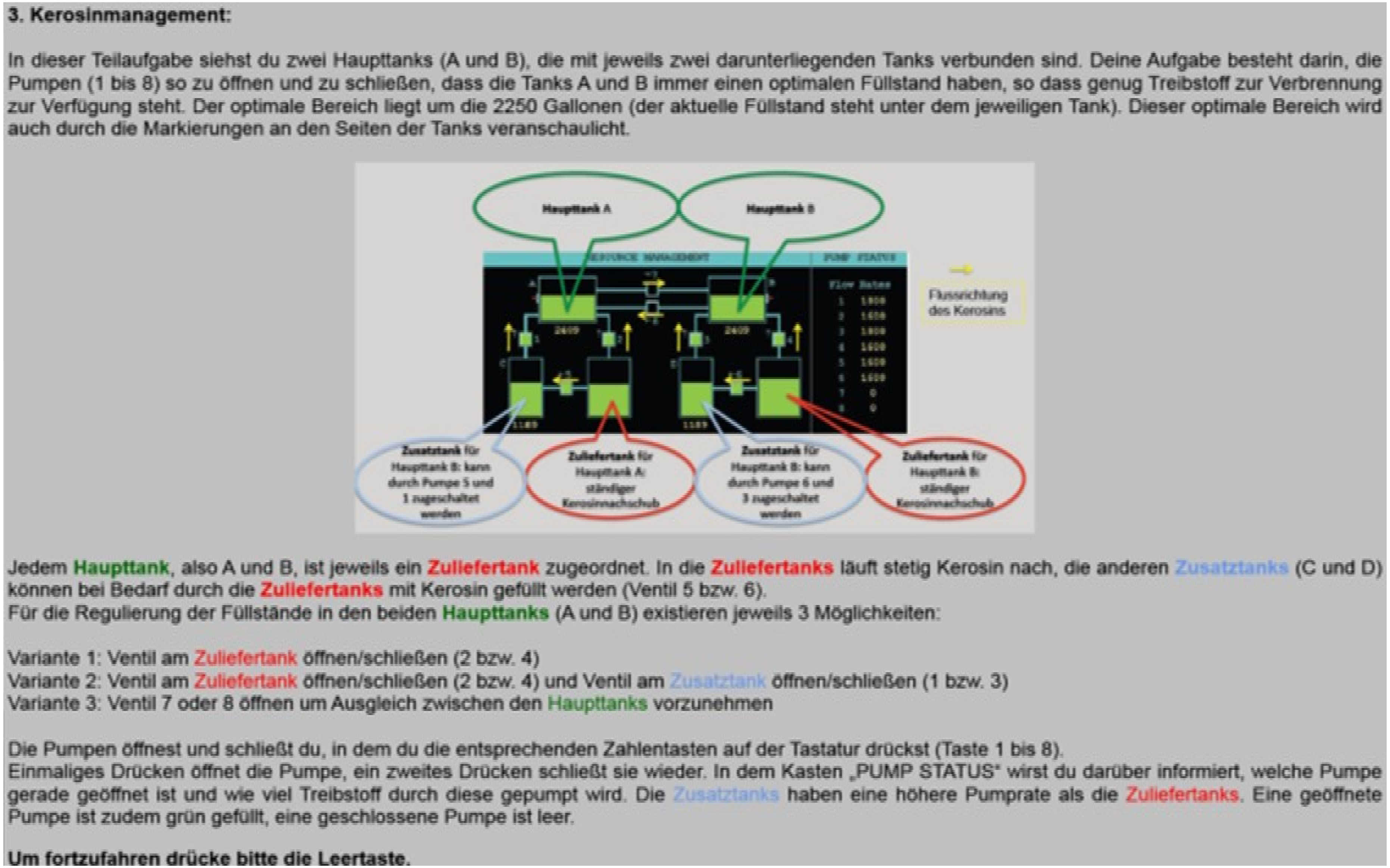
Um fortzufahren drücke bitte die Leertaste.



2. Flugzeugsteuerung:

Das Flugzeug steuerst du mit Hilfe des Joysticks. Wenn du den grünen Kreis immer in der Mitte des Fadenkreuzes halten kannst, fliegst du einen optimalen Kurs. Dabei verhält sich die Steuerung genauso wie die reale Flugzeugsteuerung: Wenn du den Joystick von dir weg drückst, sinkt das Flugzeug / der grüne Kreis. Wenn du den Joystick zu dir ran ziehst, dann steigt das Flugzeug / der grüne Kreis.
Bitte beachte, dass die Steuerung mit zunehmendem Abweichung zum optimalen Kurs schwieriger wird.

Um fortzufahren drücke bitte die Leertaste.



3. Kerosinmanagement:

In dieser Teilaufgabe siehst du zwei Haupttanks (A und B), die mit jeweils zwei darunterliegenden Tanks verbunden sind. Deine Aufgabe besteht darin, die Pumpen (1 bis 8) so zu öffnen und zu schließen, dass die Tanks A und B immer einen optimalen Füllstand haben, so dass genug Treibstoff zur Verbrennung zur Verfügung steht. Der optimale Bereich liegt um die 2250 Gallonen (der aktuelle Füllstand steht unter dem jeweiligen Tank). Dieser optimale Bereich wird auch durch die Markierungen an den Seiten der Tanks veranschaulicht.

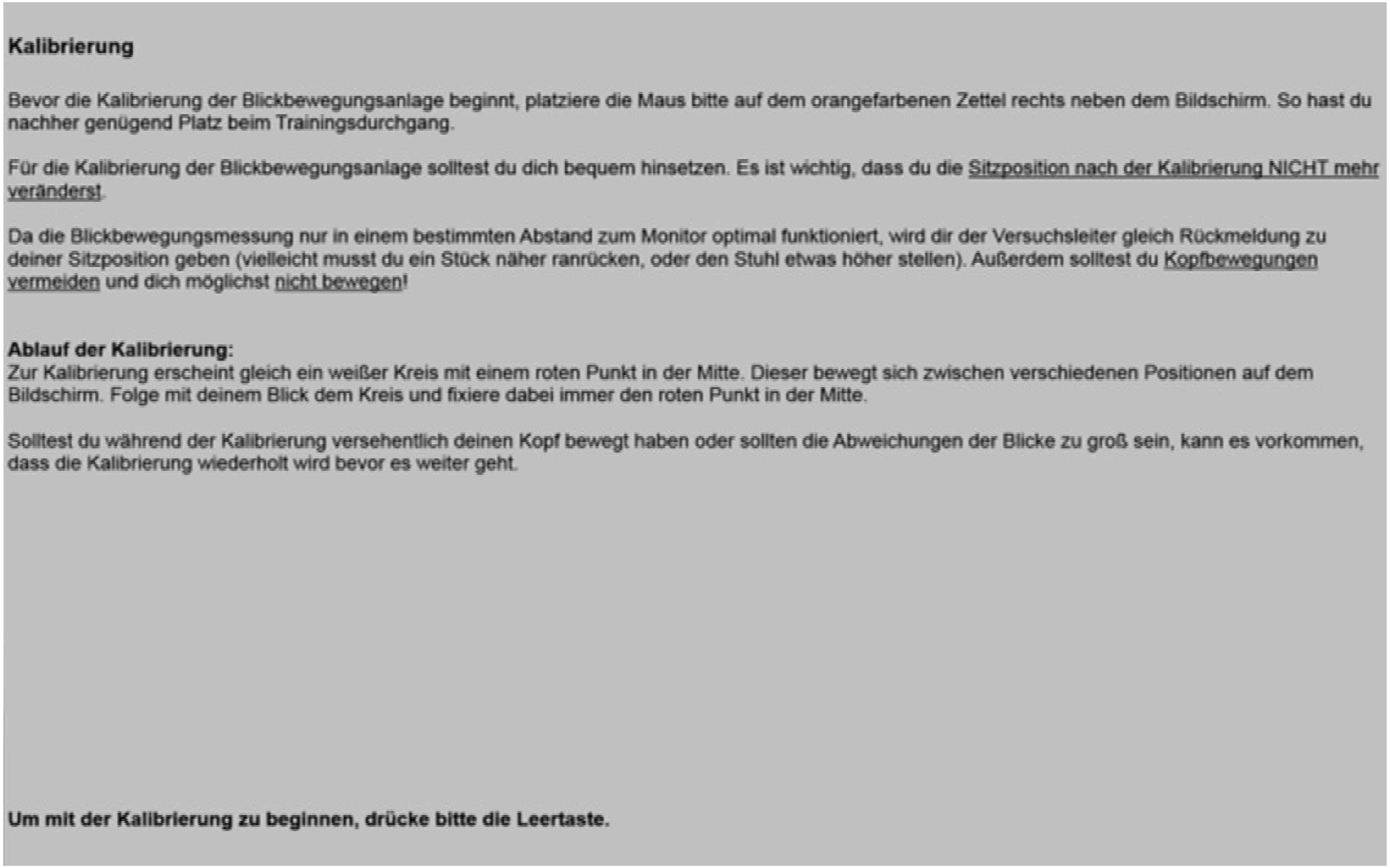
Haupttank A
Haupttank B
PUMP STATUS
Flow Rates
1 1800
2 1600
3 1800
4 1600
5 1600
6 1600
7 0
8 0
2409
2409
1189
1189
Flussrichtung des Kerosins
Zusatztank für Haupttank B: kann durch Pumpe 5 und 1 zugeschaltet werden
Zuliefertank für Haupttank A: ständiger Kerosinnachschub
Zusatztank für Haupttank B: kann durch Pumpe 6 und 3 zugeschaltet werden
Zuliefertank für Haupttank B: ständiger Kerosinnachschub

Jedem **Haupttank**, also A und B, ist jeweils ein **Zuliefertank** zugeordnet. In die **Zuliefertanks** läuft stetig Kerosin nach, die anderen Zusatztanks (C und D) können bei Bedarf durch die **Zuliefertanks** mit Kerosin gefüllt werden (Ventil 5 bzw. 6).
Für die Regulierung der Füllstände in den beiden **Haupttanks** (A und B) existieren jeweils 3 Möglichkeiten:

Variante 1: Ventil am Zuliefertank öffnen/schließen (2 bzw. 4)
Variante 2: Ventil am Zuliefertank öffnen/schließen (2 bzw. 4) und Ventil am Zusatztank öffnen/schließen (1 bzw. 3)
Variante 3: Ventil 7 oder 8 öffnen um Ausgleich zwischen den Haupttanks vorzunehmen

Die Pumpen öffnest und schließt du, in dem du die entsprechenden Zahlentasten auf der Tastatur drückst (Taste 1 bis 8).
Einmaliges Drücken öffnet die Pumpe, ein zweites Drücken schließt sie wieder. In dem Kasten „PUMP STATUS“ wirst du darüber informiert, welche Pumpe gerade geöffnet ist und wie viel Treibstoff durch diese gepumpt wird. Die Zusatztanks haben eine höhere Pumprate als die Zuliefertanks. Eine geöffnete Pumpe ist zudem grün gefüllt, eine geschlossene Pumpe ist leer.

Um fortzufahren drücke bitte die Leertaste.



Kalibrierung

Bevor die Kalibrierung der Blickbewegungsanlage beginnt, platziere die Maus bitte auf dem orangefarbenen Zettel rechts neben dem Bildschirm. So hast du nachher genügend Platz beim Trainingsdurchgang.

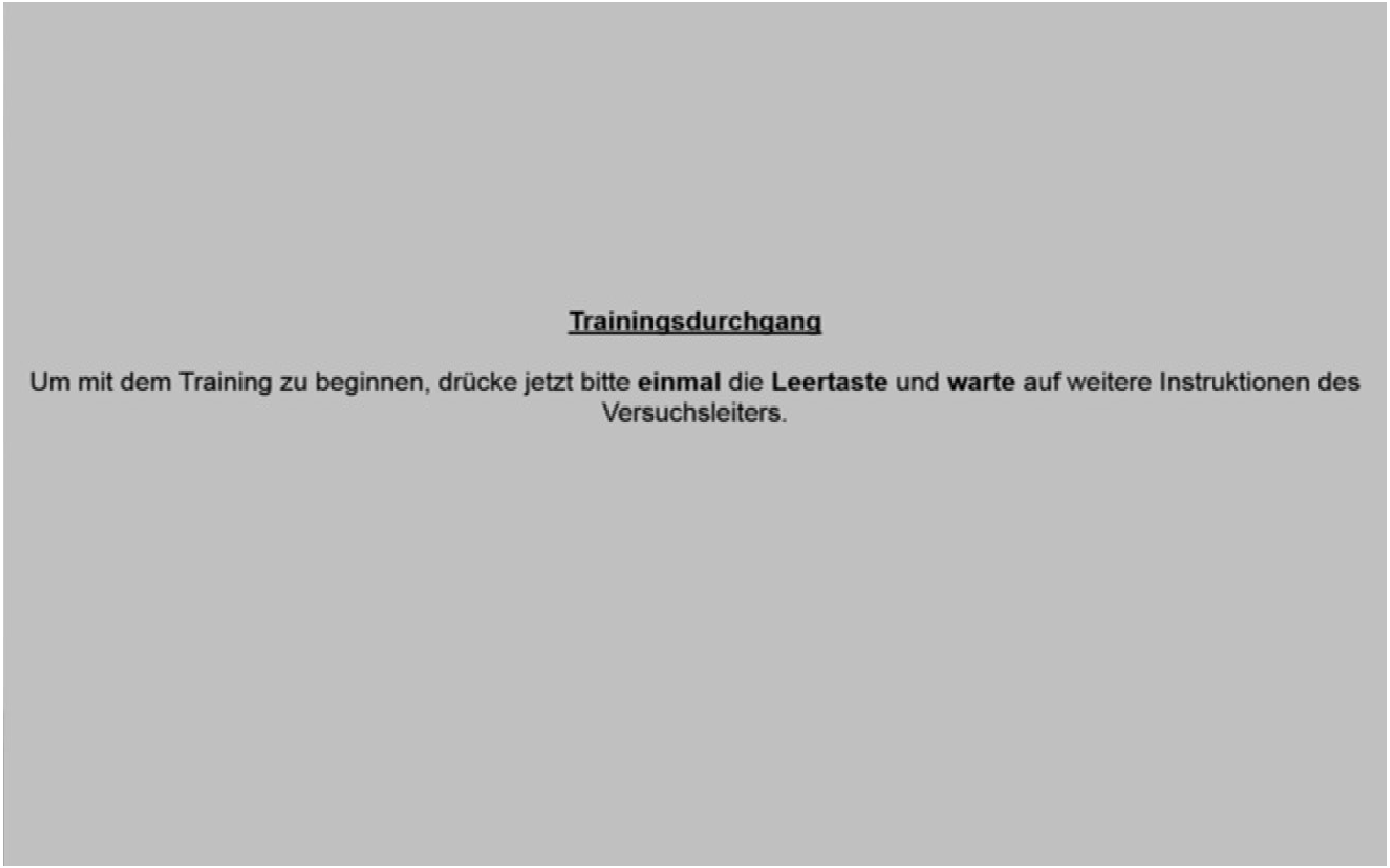
Für die Kalibrierung der Blickbewegungsanlage solltest du dich bequem hinsetzen. Es ist wichtig, dass du die Sitzposition nach der Kalibrierung NICHT mehr veränderst.

Da die Blickbewegungsmessung nur in einem bestimmten Abstand zum Monitor optimal funktioniert, wird dir der Versuchsleiter gleich Rückmeldung zu deiner Sitzposition geben (vielleicht musst du ein Stück näher ranrücken, oder den Stuhl etwas höher stellen). Außerdem solltest du Kopfbewegungen vermeiden und dich möglichst nicht bewegen!

Ablauf der Kalibrierung:
Zur Kalibrierung erscheint gleich ein weißer Kreis mit einem roten Punkt in der Mitte. Dieser bewegt sich zwischen verschiedenen Positionen auf dem Bildschirm. Folge mit deinem Blick dem Kreis und fixiere dabei immer den roten Punkt in der Mitte.

Solltest du während der Kalibrierung versehentlich deinen Kopf bewegt haben oder sollten die Abweichungen der Blicke zu groß sein, kann es vorkommen, dass die Kalibrierung wiederholt wird bevor es weiter geht.

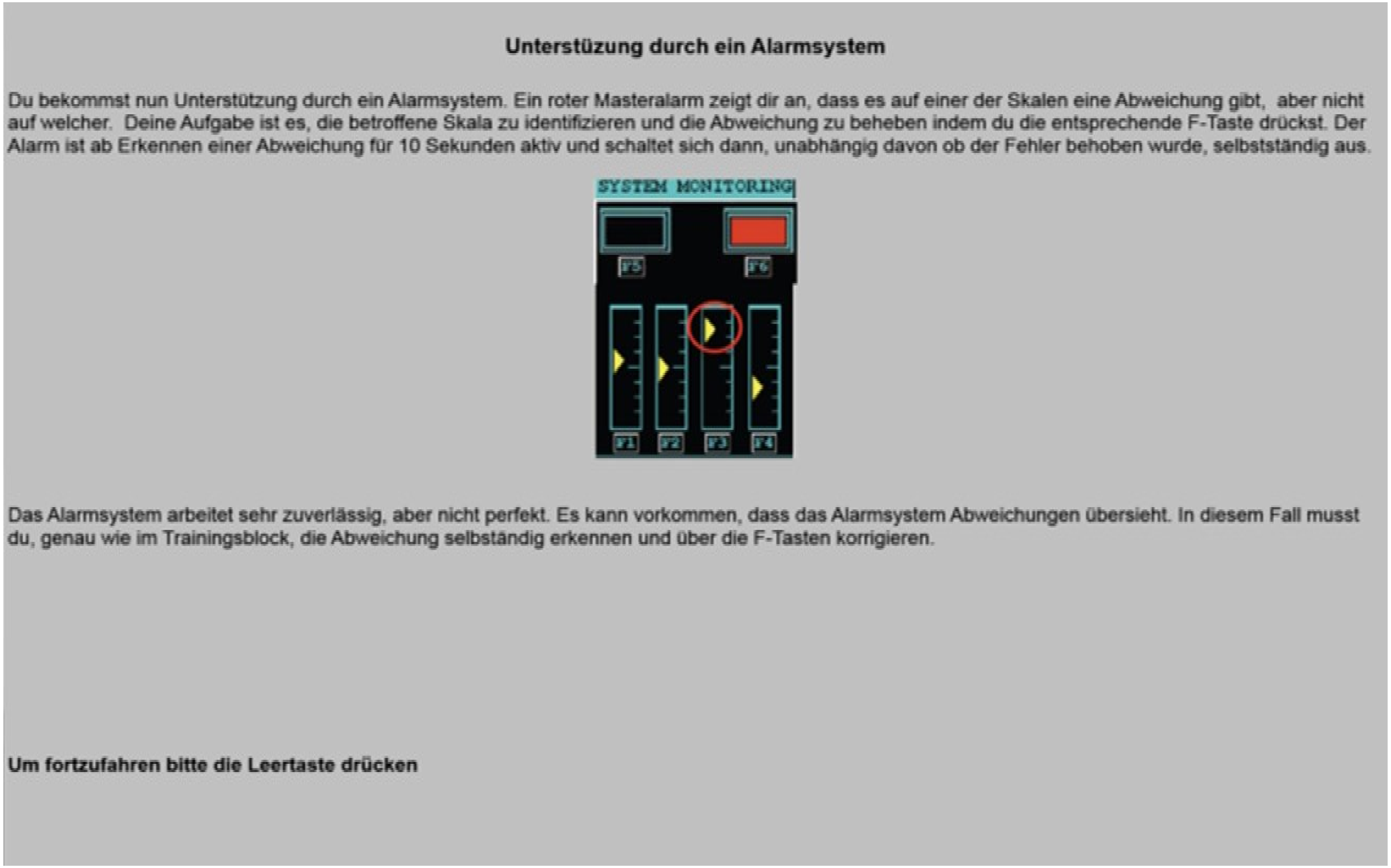
Um mit der Kalibrierung zu beginnen, drücke bitte die Leertaste.



Trainingsdurchgang

Um mit dem Training zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

This slide was only provided to the IA-supported groups.



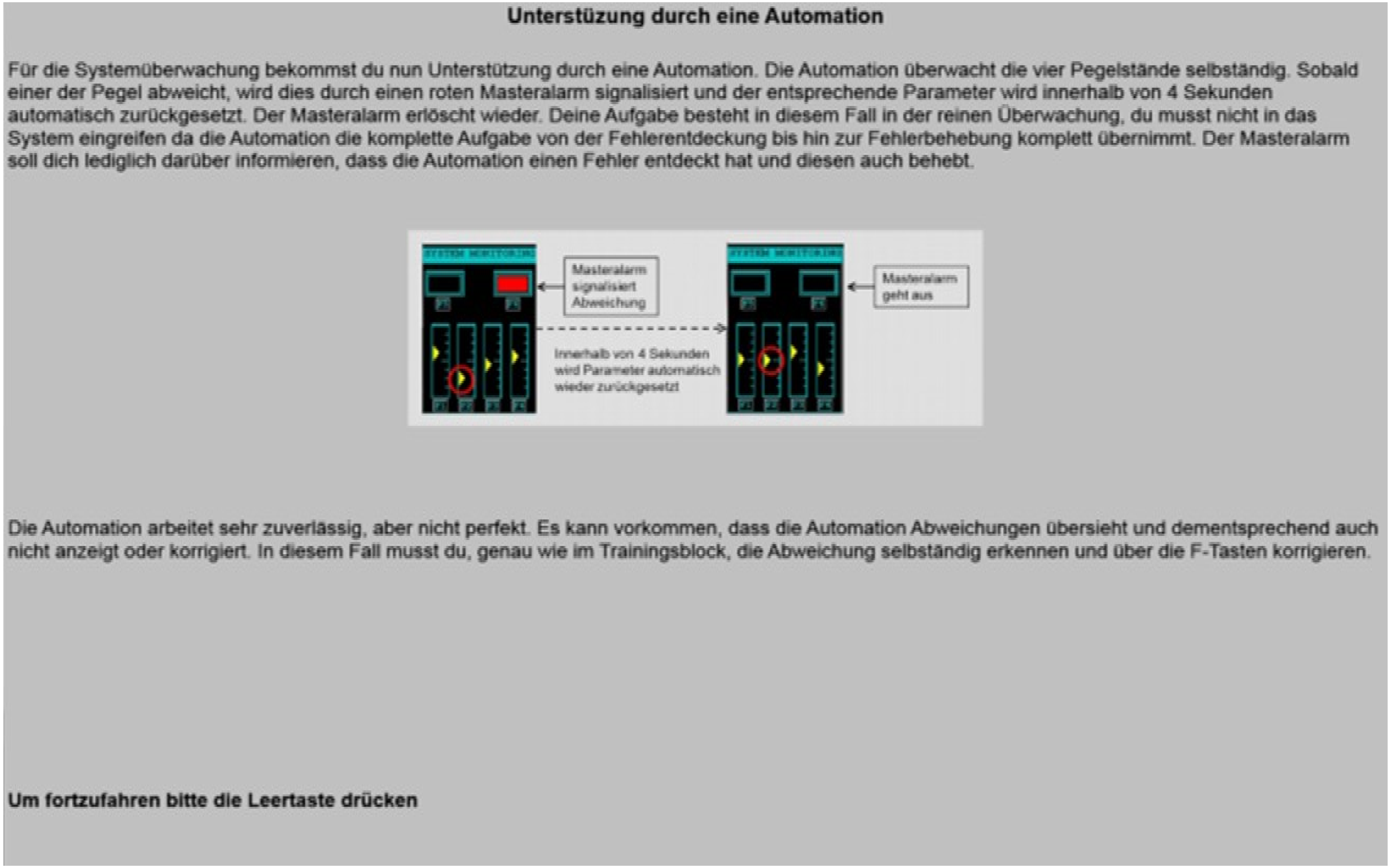
Unterstüzung durch ein Alarmsystem

Du bekommst nun Unterstützung durch ein Alarmsystem. Ein roter Masteralarm zeigt dir an, dass es auf einer der Skalen eine Abweichung gibt, aber nicht auf welcher. Deine Aufgabe ist es, die betroffene Skala zu identifizieren und die Abweichung zu beheben indem du die entsprechende F-Taste drückst. Der Alarm ist ab Erkennen einer Abweichung für 10 Sekunden aktiv und schaltet sich dann, unabhängig davon ob der Fehler behoben wurde, selbstständig aus.

Das Alarmsystem arbeitet sehr zuverlässig, aber nicht perfekt. Es kann vorkommen, dass das Alarmsystem Abweichungen übersieht. In diesem Fall musst du, genau wie im Trainingsblock, die Abweichung selbständig erkennen und über die F-Tasten korrigieren.

Um fortzufahren bitte die Leertaste drücken

This slide was only provided to the DA-supported groups.


SYSTEM MONITORING
Masteralarm signalisiert Abweichung
Innerhalb von 4 Sekunden wird Parameter automatisch wieder zurückgesetzt
SYSTEM MONITORING
Masteralarm geht aus

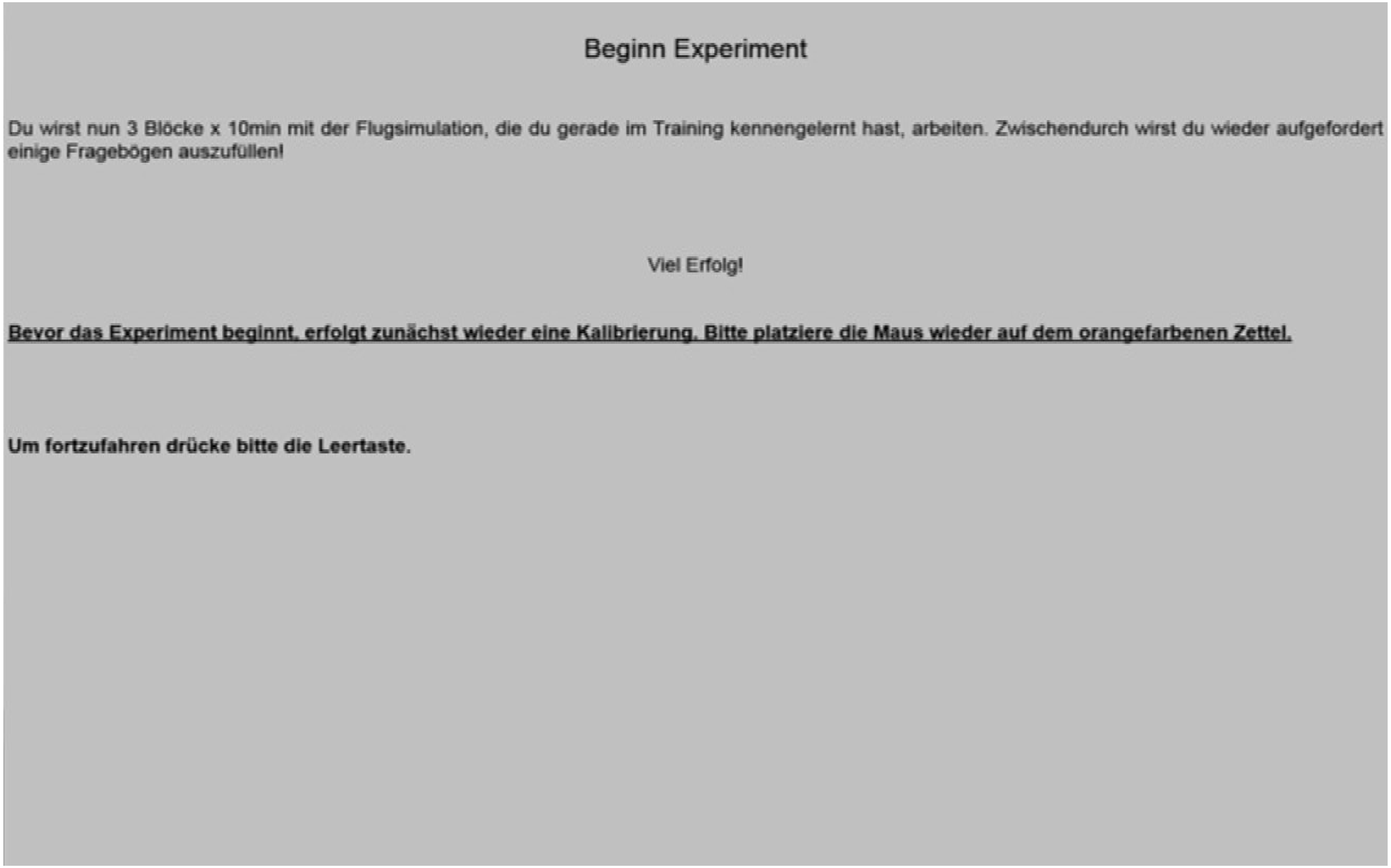
Unterstüzung durch eine Automation

Für die Systemüberwachung bekommst du nun Unterstützung durch eine Automation. Die Automation überwacht die vier Pegelstände selbständig. Sobald einer der Pegel abweicht, wird dies durch einen roten Masteralarm signalisiert und der entsprechende Parameter wird innerhalb von 4 Sekunden automatisch zurückgesetzt. Der Masteralarm erlöscht wieder. Deine Aufgabe besteht in diesem Fall in der reinen Überwachung, du musst nicht in das System eingreifen da die Automation die komplette Aufgabe von der Fehlerentdeckung bis hin zur Fehlerbehebung komplett übernimmt. Der Masteralarm soll dich lediglich darüber informieren, dass die Automation einen Fehler entdeckt hat und diesen auch behebt.

Die Automation arbeitet sehr zuverlässig, aber nicht perfekt. Es kann vorkommen, dass die Automation Abweichungen übersieht und dementsprechend auch nicht anzeigt oder korrigiert. In diesem Fall musst du, genau wie im Trainingsblock, die Abweichung selbständig erkennen und über die F-Tasten korrigieren.

Um fortzufahren bitte die Leertaste drücken

This slide was only provided to the manual control groups.



Beginn Experiment

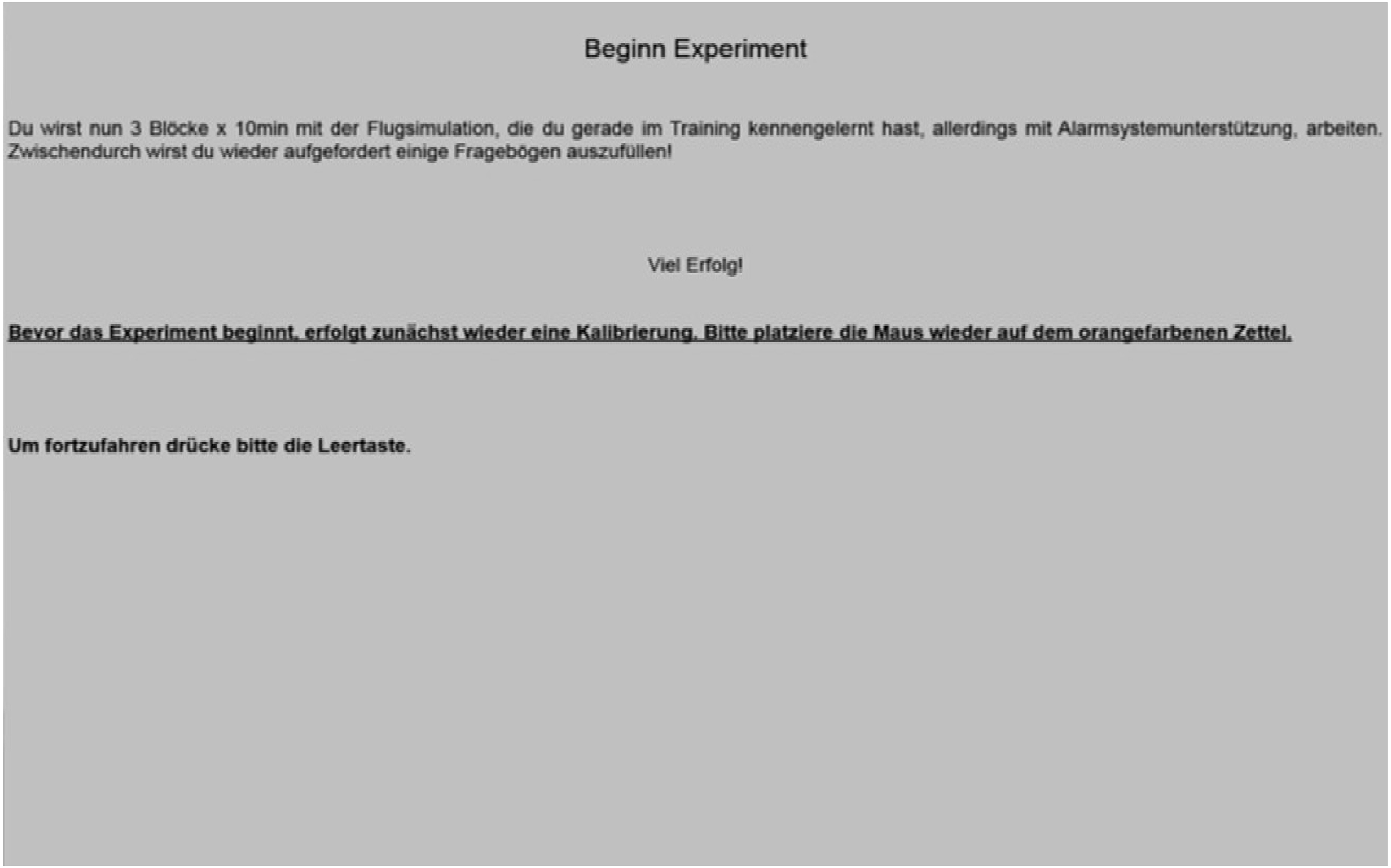
Du wirst nun 3 Blöcke x 10min mit der Flugsimulation, die du gerade im Training kennengelernt hast, arbeiten. Zwischendurch wirst du wieder aufgefordert einige Fragebögen auszufüllen!

Viel Erfolg!

Bevor das Experiment beginnt, erfolgt zunächst wieder eine Kalibrierung. Bitte platziere die Maus wieder auf dem orangefarbenen Zettel.

Um fortzufahren drücke bitte die Leertaste.

This slide was the equivalent for the IA-supported groups.



Beginn Experiment

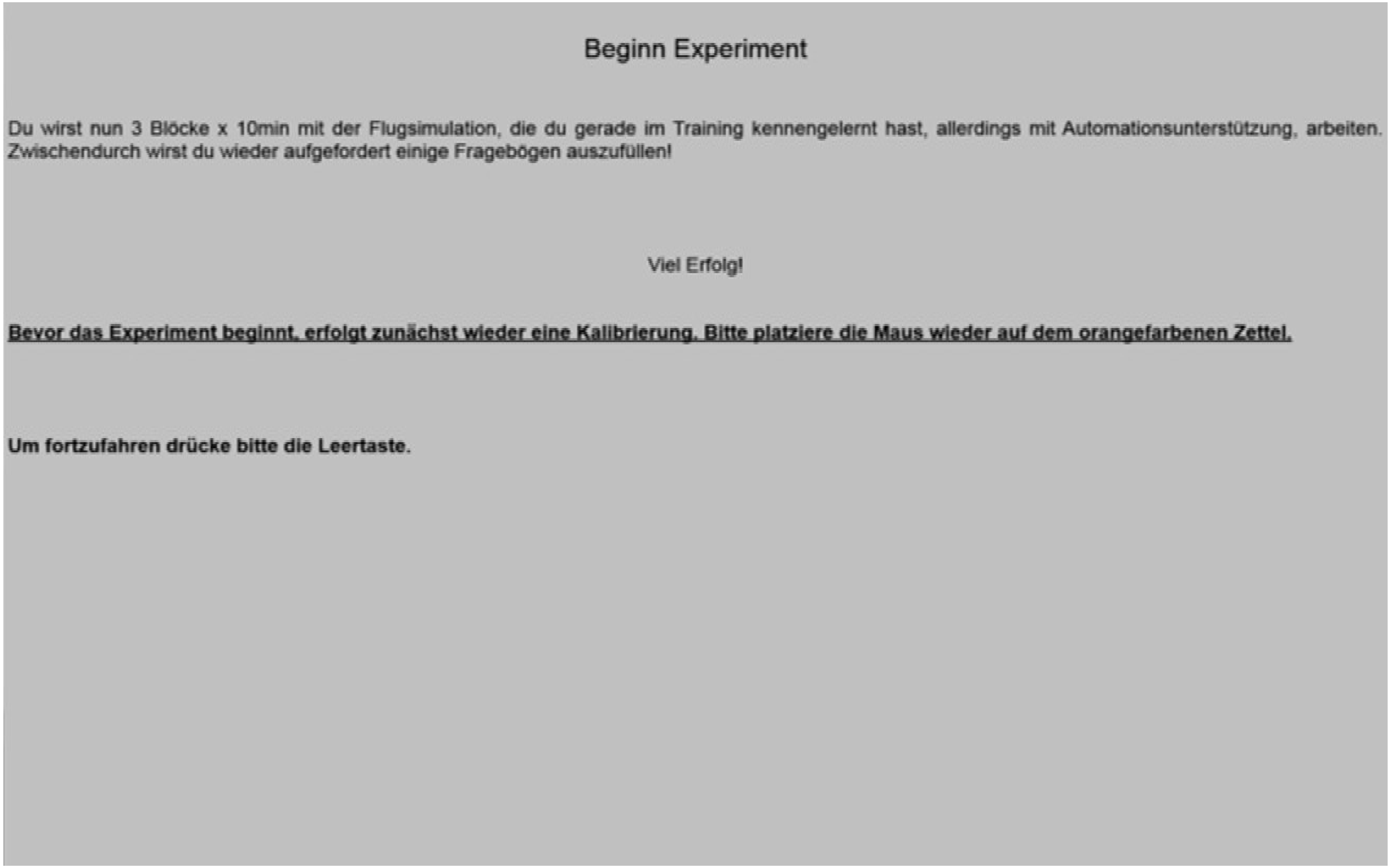
Du wirst nun 3 Blöcke x 10min mit der Flugsimulation, die du gerade im Training kennengelernt hast, allerdings mit Alarmsystemunterstützung, arbeiten. Zwischendurch wirst du wieder aufgefordert einige Fragebögen auszufüllen!

Viel Erfolg!

Bevor das Experiment beginnt, erfolgt zunächst wieder eine Kalibrierung. Bitte platziere die Maus wieder auf dem orangefarbenen Zettel.

Um fortzufahren drücke bitte die Leertaste.

This slide was the equivalent for the DA-supported groups.



Beginn Experiment

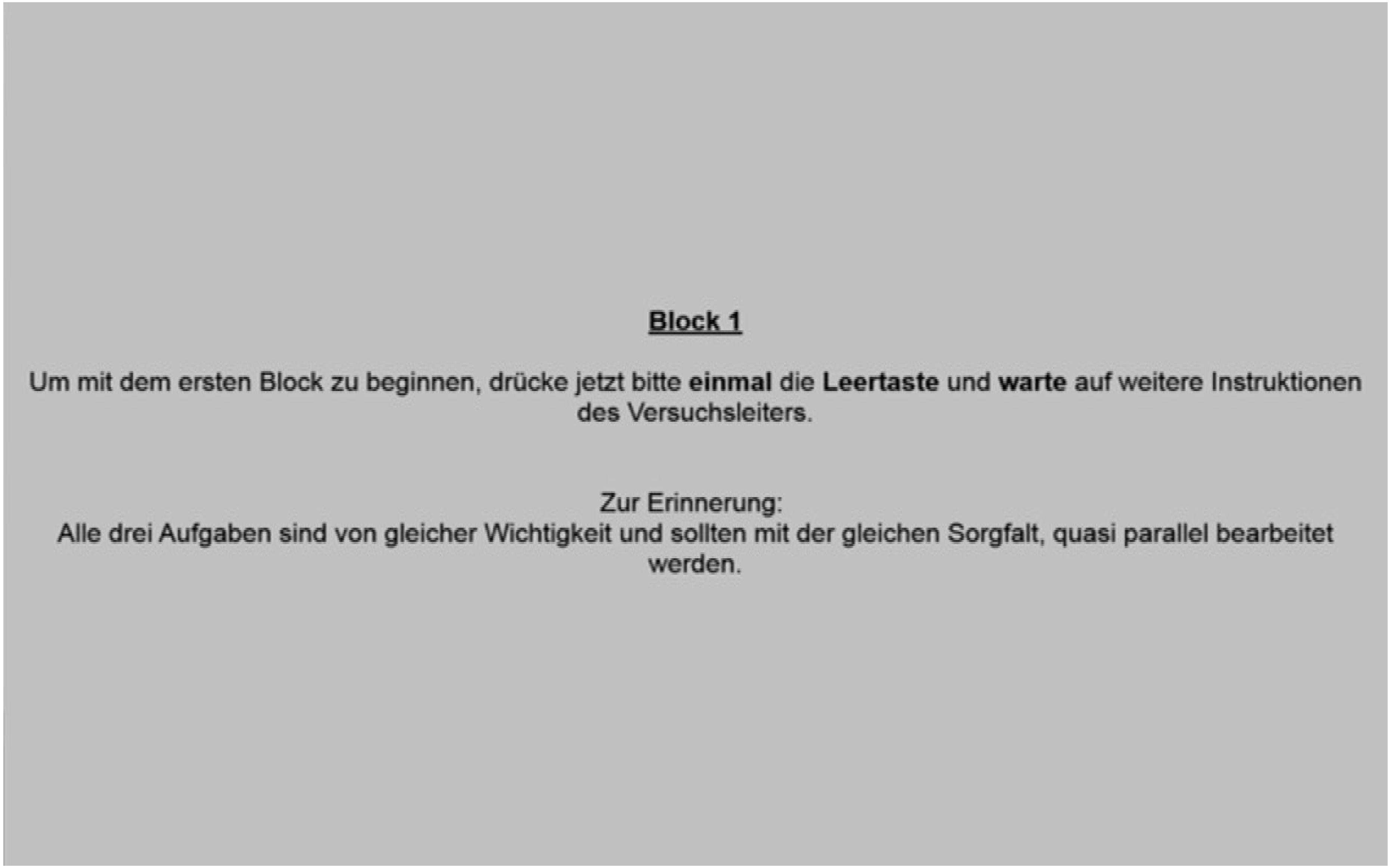
Du wirst nun 3 Blöcke x 10min mit der Flugsimulation, die du gerade im Training kennengelernt hast, allerdings mit Automationsunterstützung, arbeiten. Zwischendurch wirst du wieder aufgefordert einige Fragebögen auszufüllen!

Viel Erfolg!

Bevor das Experiment beginnt, erfolgt zunächst wieder eine Kalibrierung. Bitte platziere die Maus wieder auf dem orangefarbenen Zettel.

Um fortzufahren drücke bitte die Leertaste.

This slide was only provided to the manual control groups.

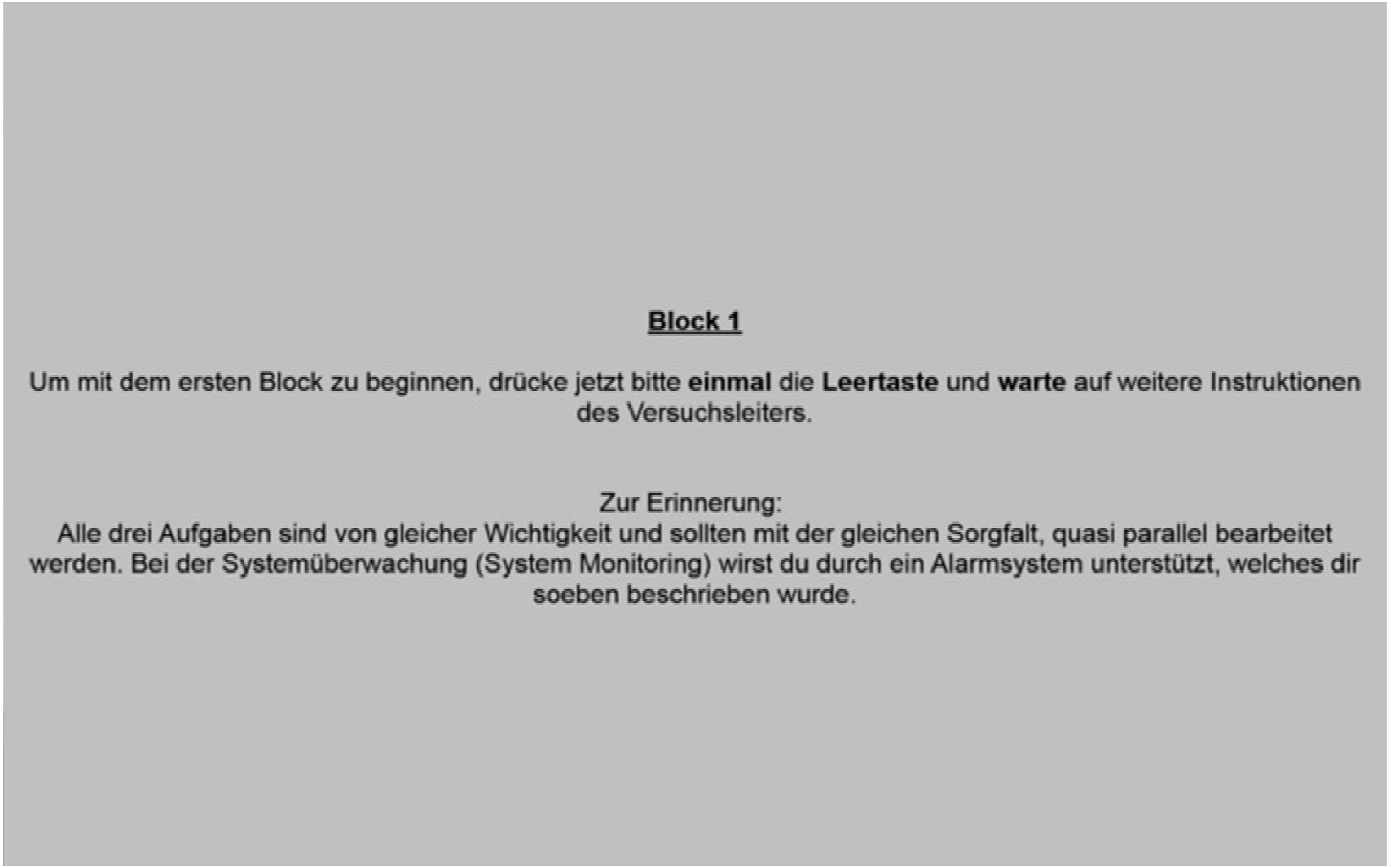


Block 1

Um mit dem ersten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.

This slide was the equivalent for the IA-supported groups.

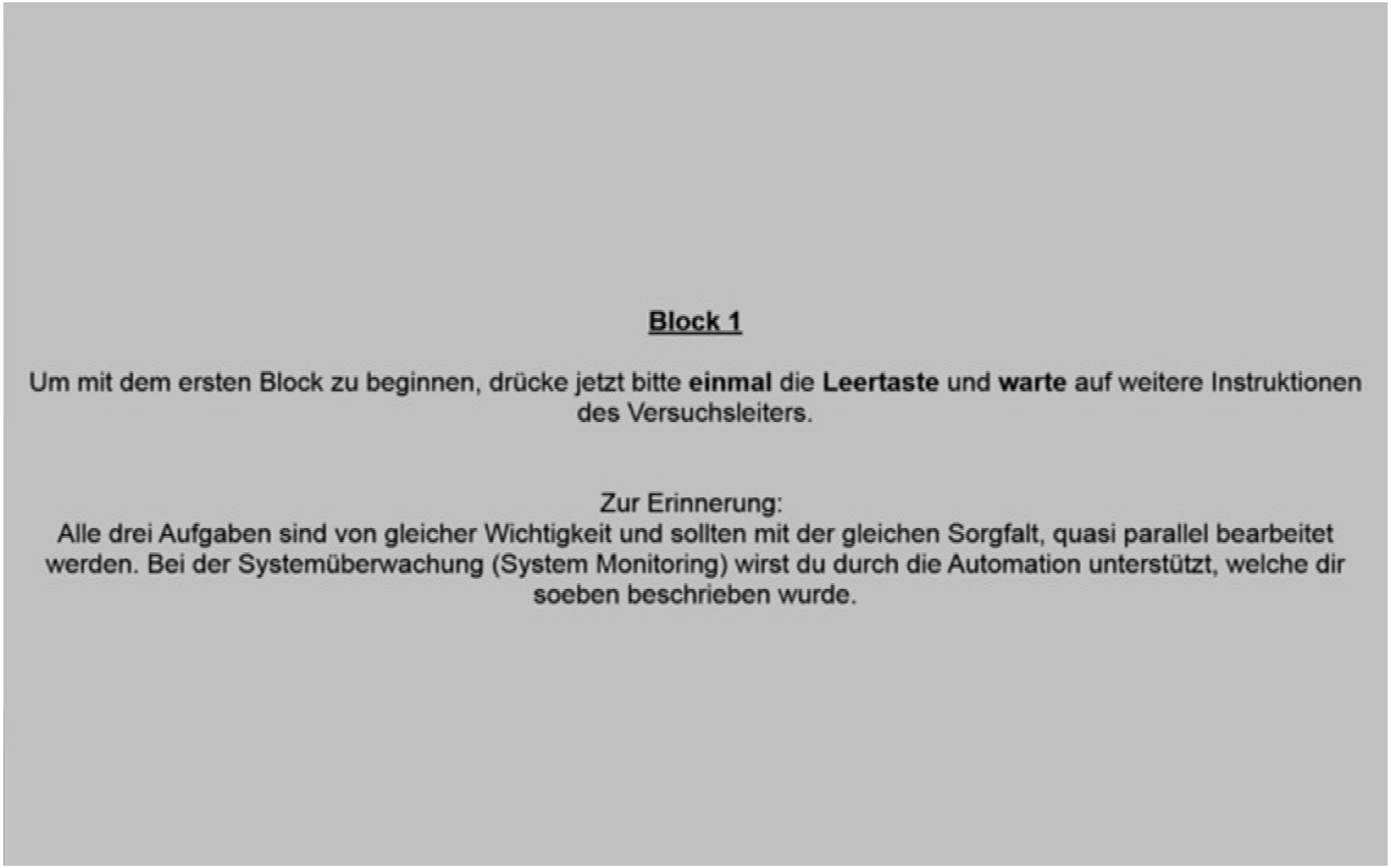


Block 1

Um mit dem ersten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden. Bei der Systemüberwachung (System Monitoring) wirst du durch ein Alarmsystem unterstützt, welches dir soeben beschrieben wurde.

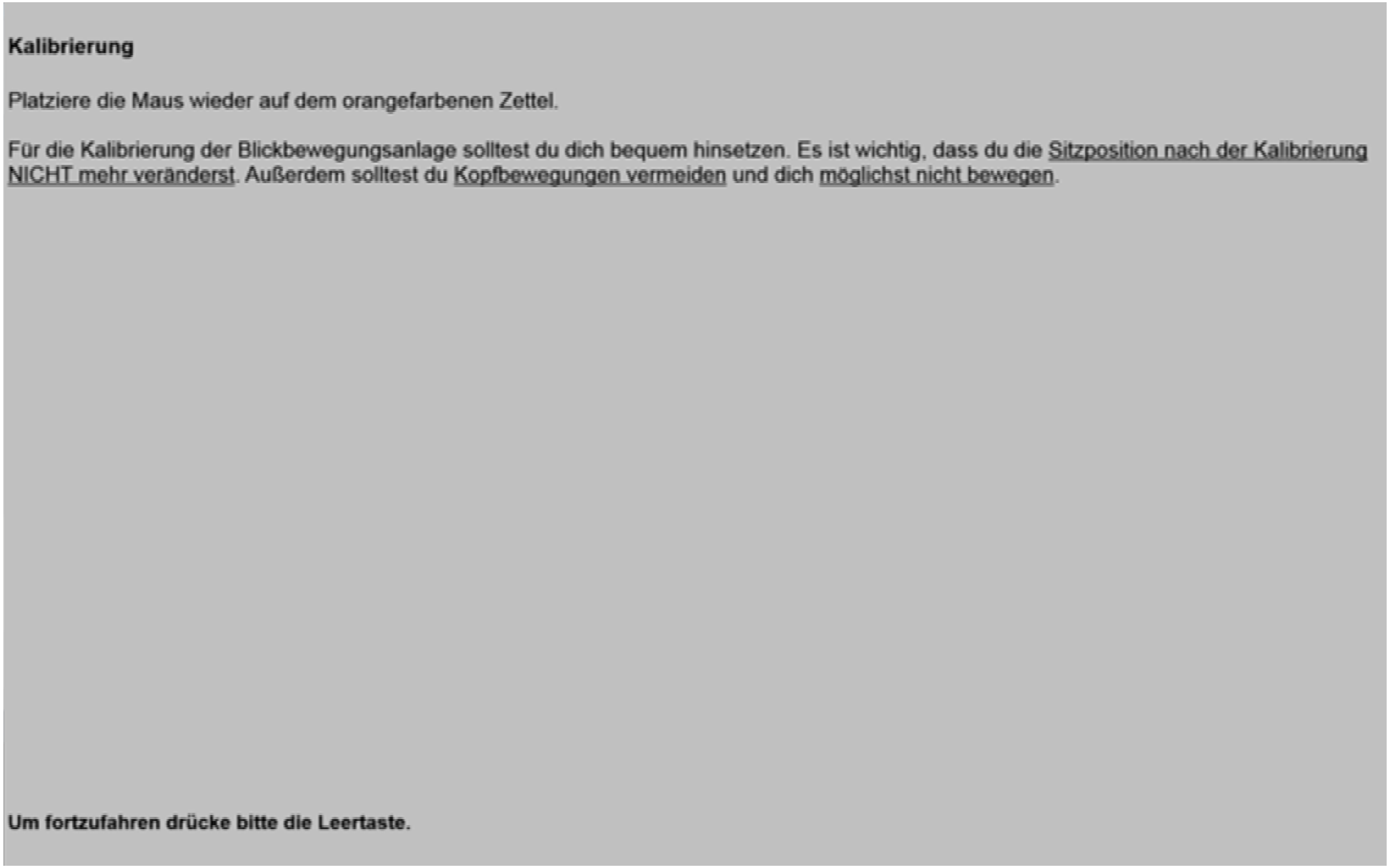
This slide was the equivalent for the DA-supported groups.



Block 1

Um mit dem ersten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden. Bei der Systemüberwachung (System Monitoring) wirst du durch die Automation unterstützt, welche dir soeben beschrieben wurde.

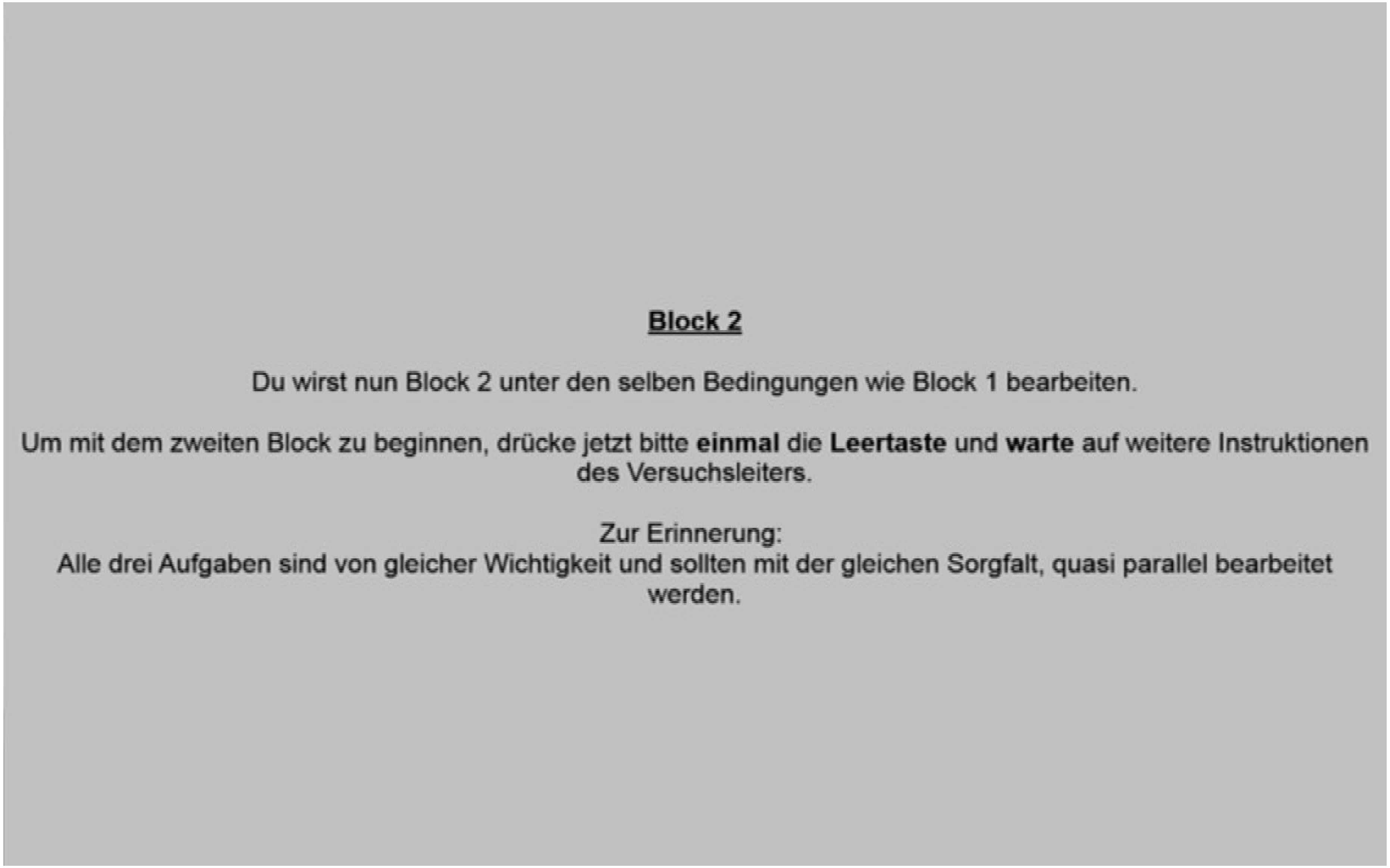


Kalibrierung

Platziere die Maus wieder auf dem orangefarbenen Zettel.

Für die Kalibrierung der Blickbewegungsanlage solltest du dich bequem hinsetzen. Es ist wichtig, dass du die Sitzposition nach der Kalibrierung NICHT mehr veränderst. Außerdem solltest du Kopfbewegungen vermeiden und dich möglichst nicht bewegen.

Um fortzufahren drücke bitte die Leertaste.

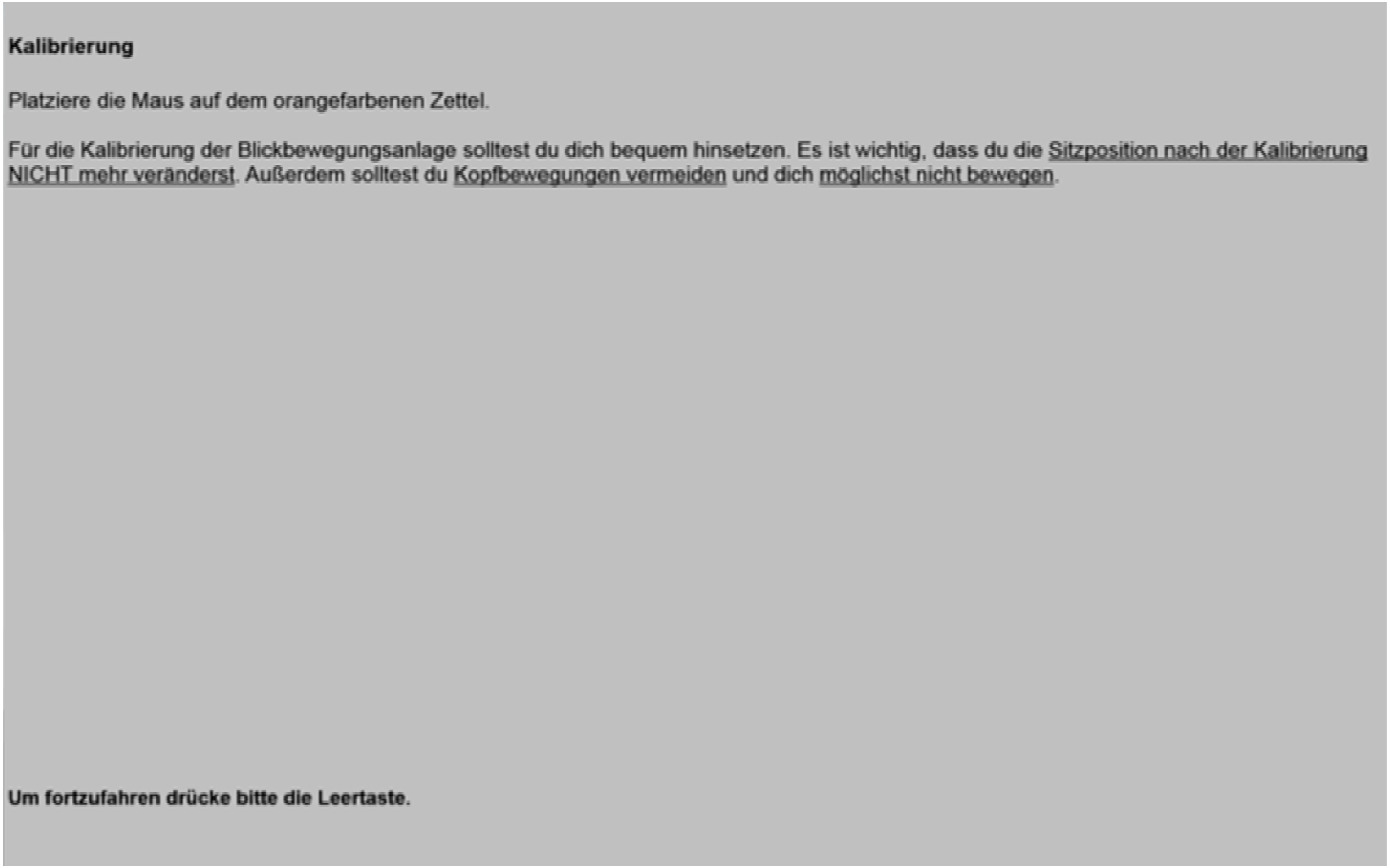


Block 2

Du wirst nun Block 2 unter den selben Bedingungen wie Block 1 bearbeiten.

Um mit dem zweiten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.

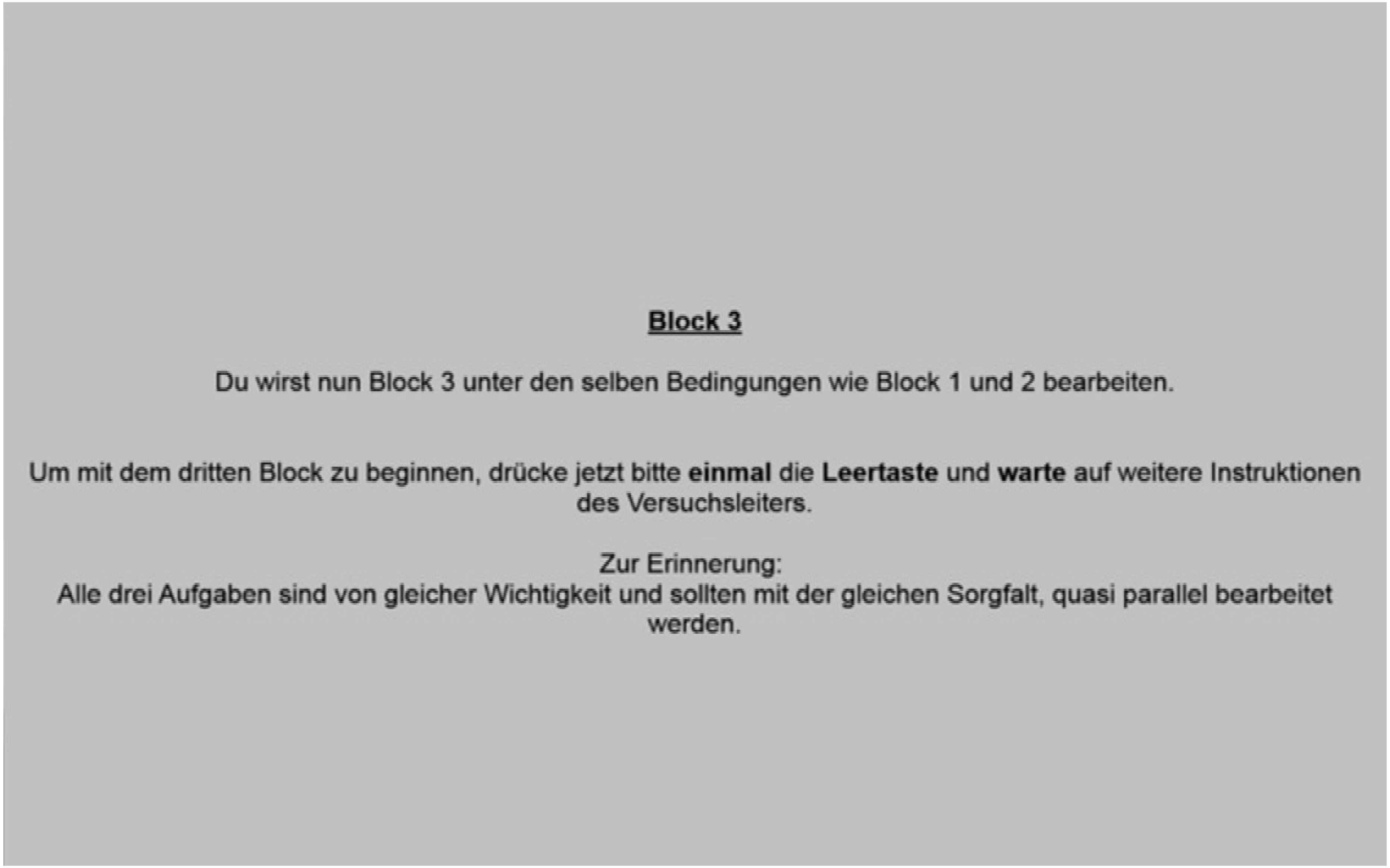


Kalibrierung

Platziere die Maus auf dem orangefarbenen Zettel.

Für die Kalibrierung der Blickbewegungsanlage solltest du dich bequem hinsetzen. Es ist wichtig, dass du die Sitzposition nach der Kalibrierung NICHT mehr veränderst. Außerdem solltest du Kopfbewegungen vermeiden und dich möglichst nicht bewegen.

Um fortzufahren drücke bitte die Leertaste.

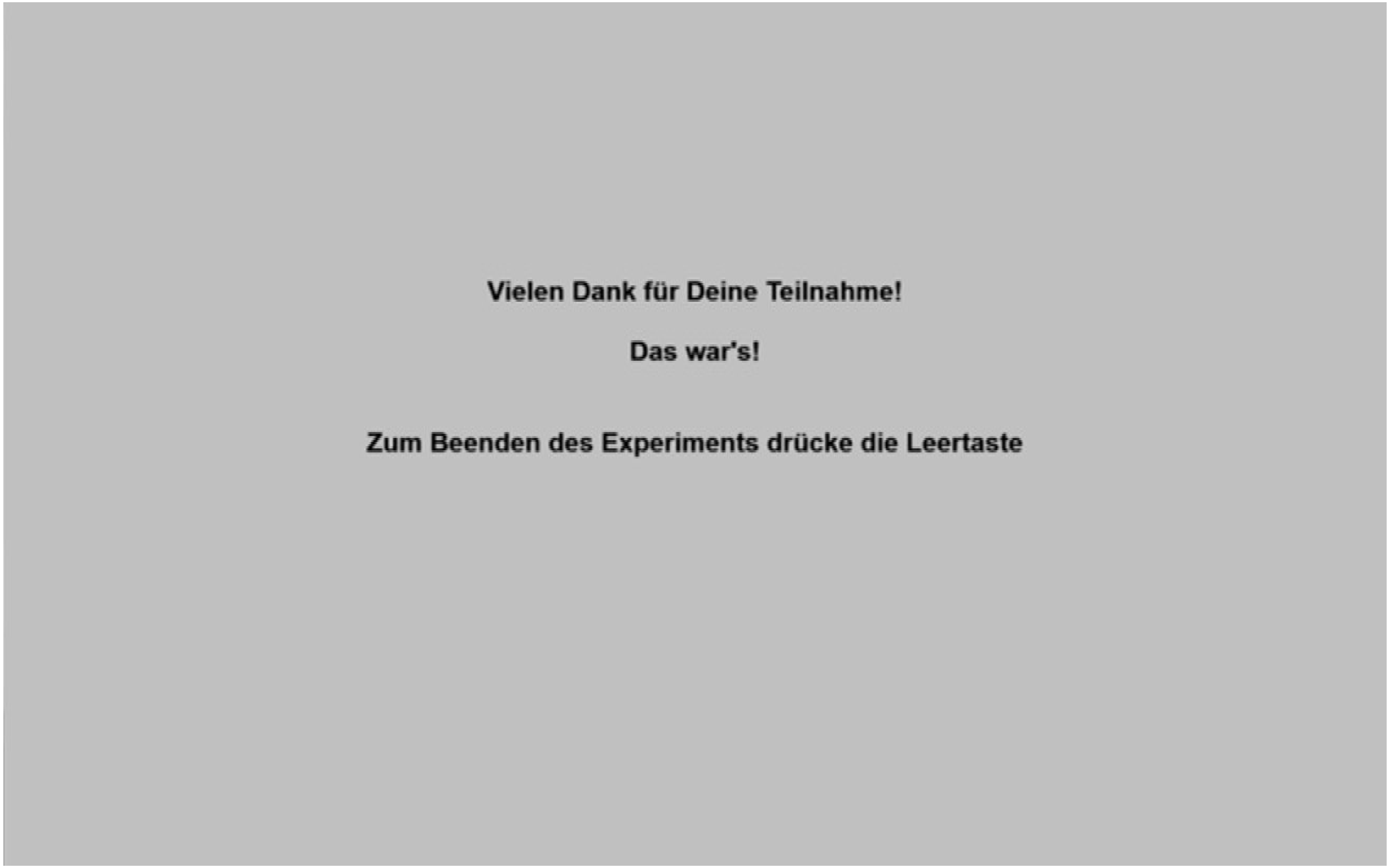


Block 3

Du wirst nun Block 3 unter den selben Bedingungen wie Block 1 und 2 bearbeiten.

Um mit dem dritten Block zu beginnen, drücke jetzt bitte **einmal** die **Leertaste** und **warte** auf weitere Instruktionen des Versuchsleiters.

Zur Erinnerung:
Alle drei Aufgaben sind von gleicher Wichtigkeit und sollten mit der gleichen Sorgfalt, quasi parallel bearbeitet werden.



Vielen Dank für Deine Teilnahme!

Das war's!

Zum Beenden des Experiments drücke die Leertaste