Technische Universität Dortmund

Fakultät Erziehungswissenschaft, Psychologie und Soziologie

Multifaktorielle Echtzeitdiagnose des Nutzerzustands in adaptiver Mensch-Maschine-Interaktion

Dissertation zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.)

vorgelegt von

Jessica Carolin Schwarz

geboren am 11.04.1983 in Stuttgart, Bad-Cannstatt

Vorgeschlagener Erstgutachter: PD Dr. Gerhard Rinkenauer

Vorgeschlagener Zweitgutachter: Prof. Dr. Josef F. Krems

März 2019

Danksagung

Diese Dissertation habe ich im Rahmen meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE) in der Abteilung Mensch-Maschine-Systeme bearbeitet. An dieser Stelle möchte ich allen Menschen danken, die mich auf dem Weg zur Promotion in den vergangenen Jahren begleitet und in unterschiedlicher Art und Weise unterstützt haben.

Besonders hervorheben möchte ich meinen Kollegen Sven Fuchs. Gemeinsam sind wir die Herausforderung angegangen, ein dynamisches adaptives System zu entwickeln und konnten uns durch unsere jeweilige thematische Fokussierung auf die Bereiche Nutzerzustandserfassung und Adaptierungsmanagement perfekt ergänzen. Viele Hürden auf dem Weg zu einem adaptiven System haben wir erfolgreich gemeistert. Ich danke für die zahlreichen hilfreichen Ratschläge, erkenntnisreichen Diskussionen und die tatkräftige Unterstützung bei den experimentellen Untersuchungen. In diesem Zusammenhang möchte ich auch Alina Schmitz-Hübsch und Andreas Werger danken; Alina insbesondere für die Unterstützung bei der Durchführung der Experimente und für das gewissenhafte Review meiner Arbeit. Andreas danke ich für die technische Implementierung der Echtzeitdiagnose RASMUS.

Für die vor allem organisatorische Unterstützung bedanke ich mich bei meiner Abteilungsleiterin Annette Kaster und meinem Forschungsgruppenleiter Oliver Witt, die mir insbesondere in der Schlussphase zeitliche Freiräume für die Fertigstellung meiner Dissertation gewährt haben, sowie bei Dr. Margarete Grandt vom BAAINBw T2.3, die das Vorhaben im Rahmen der Projekte AMISTAD und AMIGOS von Amtsseite betreut hat.

Ganz besonderer Dank gilt PD Dr. Gerhard Rinkenauer, der nach einer zufälligen Begegnung auf einer Konferenz, auf der ich einen Teil meiner Arbeit vorstellte, die Betreuung meiner Promotion als Erstgutachter übernahm. Ebenfalls danke ich Prof. Dr. Josef Krems für die Übernahme des Zweitgutachtens und die Möglichkeit, meine Arbeit im Forschungskolloquium der TU Chemnitz vorzustellen.

Fachliche Beratung und Unterstützung erfuhr ich außerdem durch Prof. Dr. Frank O. Flemisch, der mich dazu anregte, die Forschungsfrage auch einmal aus anderen Blickwinkeln, wie den Ingenieurswissenschaften zu betrachten, Dr. Daniel Feiser, der stets ein offenes Ohr für Statistik-Fragen hatte und mich dazu motivierte am Ball zu bleiben, den Kollegen des IfaDo Dr. Johanna Renker und Dr. Thorsten Plewan sowie dem Graduiertenkolleg 1855 der TU Dortmund.

Viele weitere Kollegen und Freunde haben mich unterstützt und mir geholfen, den Kopf auch einmal frei zu bekommen. Stellvertretend nennen möchte ich Daniel Schimikowski, Christina Seimetz, Lerke Thiele sowie meinen neuen Teamkollegen Thomas Witte, mit dem ich an diesem spannenden Forschungsfeld weiterarbeiten werde.

Zu guter Letzt geht mein herzlichster Dank an meine Eltern Michael und Margret Schwarz, die jederzeit für mich da sind, und denen ich so viel zu verdanken habe!

Zusammenfassung

Konzepte adaptiver Systemgestaltung zielen darauf ab, den Operateur zustands- und situationsabhängig zu unterstützen, um Leistungsausfällen und -minderungen frühzeitig entgegenzuwirken und die Gesamtleistung des Mensch-Maschine-Systems zu optimieren. Gegenstand dieser Dissertation ist die Entwicklung und Untersuchung einer technikseitigen Nutzerzustandsdiagnose, die eine ganzheitliche Erfassung und Bewertung relevanter mentaler Zustände und Einflussfaktoren bereits während der Aufgabenbearbeitung ermöglicht und damit die Grundlage für die Auswahl und Anwendung von Unterstützungsstrategien bildet.

In vergangenen Studien zur Nutzerzustandsdiagnose in adaptiven Systemen zeigte sich, dass die Umsetzung einer solchen Diagnosefunktion insbesondere für Anwendungen in der realen Welt mit Herausforderungen verbunden ist. Eine Herausforderung besteht darin, dass außerhalb von kontrollierten Laborumgebungen eine Vielzahl von Faktoren auf das Mensch-Maschine-System einwirkt, die den Nutzerzustand und die menschliche Leistungsfähigkeit beeinflussen. Daher greift eine einseitige Betrachtung des Nutzerzustands und ein symptomatisches Reagieren auf Zustandsveränderungen oft zu kurz (vgl. Steinhauser et al., 2009). Es zeigte sich auch, dass einzelne Diagnosemaße, wie die Herzrate oder die Pupillenweite, nicht ausreichen, um den Nutzerzustand verlässlich zu diagnostizieren (Grandt, 2004, Veltman & Jansen, 2006). Bei Verwendung von subjektiven, physiologischen und verhaltensbasierten Diagnosemaßen bleiben zudem die Ursachen für Zustandsveränderungen unklar, die für kontextspezifische und problemadäquate Adaptierungen grundlegend sind.

Um diesen Herausforderungen zu begegnen, wurde im Rahmen der Promotion ein Konzept für eine multifaktorielle Bewertung des Nutzerzustands entwickelt. Der Nutzerzustand wird dabei als Zusammenwirken von sechs verschiedenen mentalen Konstrukten definiert, die nachweislich die menschliche Leistungsfähigkeit positiv oder negativ beeinflussen können. Neben der in vielen Studien betrachteten mentalen Beanspruchung betrifft dies die Müdigkeit, die Motivation, die Aufmerksamkeit, das Situationsbewusstsein sowie den emotionalen Zustand.

Der Ansatz der multifaktoriellen Nutzerzustandsbewertung sieht des Weiteren vor, neben physiologischen und verhaltensbasierten Diagnosemaßen auch umwelt- und nutzerinterne Einflussfaktoren in die Diagnose einzubeziehen, um Rückschlüsse auf die Ursachen für Nutzerzustandsveränderungen zu ermöglichen. Auf Basis von psychologischen Modellen und Erkenntnissen zum Nutzerzustand wurde ein generisches Modell zum Nutzerzustand erstellt, das aufzeigt, welche Einflussfaktoren auf die menschliche Leistungsfähigkeit einwirken, die in der Diagnose zu berücksichtigen sind, um zielgerichtete Adaptierungen zu ermöglichen.

Die theoretischen Erkenntnisse wurden in zwei experimentellen Untersuchungen überprüft. In Experiment 1 (N=12) wurden Eyetracking-Maße sowie Maße eines EEG herangezogen, um die Diagnosefähigkeit dieser Maße hinsichtlich der Nutzerzustandsdimensionen Beanspruchung, Aufmerksamkeit und Frustration zu untersuchen. Des Weiteren wurde in diesem Experiment der Einfluss von umwelt- und nutzerinternen Faktoren untersucht. Ein Jahr später wurde ein Retest mit 10 Probanden aus dem ersten Experiment durchgeführt, um die zeitliche Stabilität der Diagnosemaße zu überprüfen (Experiment 2).

Auf Basis der Erkenntnisse aus der theoretischen Analyse und den beiden empirischen Untersuchungen wurde *RASMUS (Real time Assessment of Multidimensional User State)* entwickelt. Dabei handelt es sich um ein generisches Diagnosekonzept für eine multifaktorielle Echtzeiterfassung und -bewertung von Nutzerzuständen, das vier Diagnoseschritte umfasst:

1. Zusammenführung und Synchronisation der zur Diagnose verwendeten Daten,
2. Bestimmung von Adaptierungsbedarf über Leistungsmaße,
3. Ursachenanalyse (Diagnose von kritischen Nutzerzuständen und Indikatoren),
4. Bereitstellung der Diagnoseergebnisse für die Auswahl von Adaptierungsstrategien.

Ein wesentliches Merkmal des Diagnosekonzepts besteht darin, dass der Adaptierungsbedarf nicht über den Nutzerzustand sondern über die Leistung bzw. das Auftreten eines Leistungseinbruchs bestimmt wird. Auf diese Weise wird sichergestellt, dass die Adaptierung produktiven Selbstregulierungsstrategien des Nutzers nicht entgegenwirkt. Bei Auftreten eines Leistungseinbruchs erfolgt sodann eine Diagnose kritisch ausgeprägter Nutzerzustände und Einflussfaktoren, die als Ursachen für den Leistungseinbruch in Betracht kommen. Die Diagnose kritischer Nutzerzustände soll es dem adaptiven System ermöglichen, Adaptierungsstrategien auszuwählen, die den kritischen Zuständen entgegenwirken und die Leistung des Nutzers wiederherstellen.

RASMUS wurde exemplarisch für den Anwendungsbereich der Luftraumüberwachung und drei in diesem Bereich besonders relevante Problemzustände (hohe Beanspruchung, passive aufgabenbezogene Müdigkeit und falsche Aufmerksamkeitsverteilung) umgesetzt und validiert. Dies erfolgte auf Basis von zwei weiteren experimentellen Untersuchungen: Experiment 3 diente dazu, geeignete Indikatoren auszuwählen und Diagnoseregeln für kritische Indikatorausprägungen und Nutzerzustände für die betrachtete Experimentalaufgabe zu definieren. In Experiment 4 wurde die Validität der Diagnoseergebnisse von RASMUS in Hinblick auf die drei genannten Problemzustände empirisch geprüft. Zur Validierung wurden die Ergebnisse von RASMUS zum Zeitpunkt von Leistungseinbrüchen mit den Ergebnissen einer subjektiven Bewertung des Nutzerzustands durch die Teilnehmer verglichen. Bei der falschen Aufmerksamkeitsverteilung wurde neben der subjektiven Bewertung auch das über eine SAGAT-Frage erfasste Level-1-Situationsbewusstsein (Level-1-SA) als Vergleichsmaß herangezogen, da dies eng mit der Aufmerksamkeit verknüpft ist und eine objektive Bewertung ermöglicht (vgl. Endsley, 1988, 1995).

In Bezug auf die detektierten Problemzustände *hohe Beanspruchung* und *passive aufgabenbezogene Müdigkeit* zeigten sich signifikante Übereinstimmungen mit den subjektiven Bewertungen. In Hinblick auf den Problemzustand *falsche Aufmerksamkeitsverteilung* konnte eine signifikante Übereinstimmung nur für das Vergleichsmaß Level-1-SA nachgewiesen werden. Die nicht erwartungskonformen Ergebnisse für die subjektive Bewertung sprechen dafür, dass die Teilnehmer Probleme hatten, ihre Aufmerksamkeit richtig einzuschätzen. Insgesamt bestätigen die Ergebnisse eine valide Erfassung der drei untersuchten Problemzustände durch RASMUS. Die erfolgreiche Umsetzung des Diagnosekonzepts und die bestätigte Validität der Diagnoseergebnisse ermöglichen es, RASMUS künftig als Grundlage für die dynamische Auswahl und Konfiguration von Unterstützungsstrategien in adaptiven Systemen heranzuziehen. Darüber hinaus sind auch Einsatzmöglichkeiten im Bereich Training und Ausbildung denkbar, z.B. um gewünschte Trainingszustände gezielt hervorzurufen oder den Trainingserfolg zu bewerten.

Inhaltsverzeichnis

1 Einführung 1

1.1 Flug AF447 1
1.1.1 Einflussfaktoren auf den Nutzerzustand 1
1.1.2 Auswirkungen auf den emotionalen Zustand 2
1.1.3 Auswirkungen auf die Beanspruchung 2
1.1.4 Auswirkungen auf das Situationsbewusstsein 2
1.1.5 Auswirkungen auf die Aufmerksamkeit 3
1.1.6 Resümee 3
1.2 Probleme bei hochautomatisierten Systemen 3
1.2.1 Aufrechterhaltung des Situationsbewusstseins 4
1.2.2 Vertrauen in die Automation 5
1.2.3 Fertigkeitsverlust 5
1.3 Lösungsansatz: adaptive Gestaltung der Mensch-Mensch-Maschine-Interaktion (MMI) 6
1.4 Eigenes Rahmenwerk adaptiver MMI 6
1.5 Zielsetzung des Promotionsvorhabens 8
1.5.1 Anforderungen an das Diagnosekonzept 8
1.5.2 Forschungsfragen 9
1.5.3 Validierung des Diagnosekonzepts 10
1.6 Vorgehensweise und Gliederung der Dissertation 10

2 Stand der Forschung 13

2.1 Bisherige Ansätze zur adaptiven Gestaltung der Mensch-Maschine-Interaktion 13
2.1.1 Adaptive Aiding 13
2.1.2 Adaptive Automation 14
2.1.3 Augmented Cognition 15
2.1.4 Resümee 15
2.2 Nutzerzustandserfassung in adaptiven Systemen 16
2.2.1 Erfassung mentaler Beanspruchung 17
2.2.2 Erfassung des emotionalen Zustands 18
2.2.3 Erfassung der Motivation 19
2.2.4 Erfassung von Müdigkeit 20
2.2.5 Erfassung der Aufmerksamkeit und Vigilanz 20
2.2.6 Erfassung des Situationsbewusstseins 21
2.2.7 Resümee 22
2.3 Bewertung der Methoden zur Nutzerzustandserfassung 22
2.3.1 Bewertungskriterien 23
2.3.2 Bewertung der Methodengruppen 23
2.3.3 Detailbewertung physiologischer und verhaltensbasierter Maße 26
2.3.4 Resümee 31
2.4 Erkenntnisse und Anforderungen für die Konzeption einer Echtzeitdiagnose 33
2.4.1 Störeinflüsse 34
2.4.2 Reliabilität und zeitliche Stabilität 34
2.4.3 Interindividuelle Unterschiede 34
2.4.4 Menschliche Selbstregulierung 35

2.4.5 Ursachenanalyse ... 35
2.4.6 Kontextfaktoren ... 36
2.4.7 Oszillieren der Adaptierung ... 36
2.4.8 Künstlichkeit von Laboraufgaben ... 37

3 Theoretische Grundlagen für eine multifaktorielle Nutzerzustandsdiagnose ... 39

3.1 Theorien und Modelle zu den sechs Dimensionen des Nutzerzustands ... 40
3.1.1 Mentale Beanspruchung ... 40
3.1.2 Emotionaler Zustand ... 45
3.1.3 Motivation ... 49
3.1.4 Müdigkeit ... 51
3.1.5 Aufmerksamkeit ... 54
3.1.6 Situationsbewusstsein ... 57
3.2 Wechselwirkungen zwischen den Nutzerzustandsdimensionen ... 59
3.3 Einflussfaktoren auf Nutzerzustand und Informationsverarbeitung ... 60
3.3.1 Individuelle Faktoren ... 61
3.3.2 Aufgaben ... 62
3.3.3 Umgebung ... 62
3.3.4 Kontext ... 63
3.3.5 Eigenschaften des technischen Systems ... 63
3.3.6 Ziele ... 64
3.3.7 Ereignisse ... 64
3.4 Ableitung eines generischen Modells zum Nutzerzustand ... 65
3.4.1 Komponenten des Modells ... 65
3.4.2 Der Informationsverarbeitungsprozess ... 66
3.4.3 Der Zustandsregulierungsprozess ... 66
3.5 Resümee ... 67

4 Experiment 1 – Untersuchung einer multifaktoriellen Nutzerzustandsbewertung ... 69

4.1 Forschungsziele ... 69
4.2 Methodisches Vorgehen ... 70
4.2.1 Experimentalaufgabe ... 70
4.2.2 Versuchsdesign ... 73
4.2.3 Unabhängige und abhängige Variablen ... 75
4.2.4 Individuelle Faktoren ... 76
4.2.5 Physiologische und verhaltensbasierte Maße ... 79
4.2.6 Subjektive Maße und Leistungsmaße ... 81
4.2.7 Hypothesen ... 82
4.2.8 Stichprobe und Versuchsdurchführung ... 86
4.2.9 Vorgehen bei der Datenaufbereitung ... 88
4.2.10 Statistische Auswertung ... 90
4.3 Ergebnisse der experimentellen Untersuchung ... 91
4.3.1 Einfluss der Anforderungssituation auf den diagnostizierten Nutzerzustand (Hypothese 1) ... 91
4.3.2 Einfluss individueller Faktoren auf den Nutzerzustand und auf die Leistung (Hypothese 2) ... 96
4.3.3 Zusammenhänge zwischen Diagnosemaßen des Nutzerzustands (Hypothese 3) ... 96
4.4 Diskussion ... 102

4.4.1 Diskussion der Ergebnisse zu Hypothese 1 ... 102
4.4.2 Diskussion der Ergebnisse zu Hypothese 2 ... 104
4.4.3 Diskussion der Ergebnisse zu Hypothese 3 ... 105
4.5 Resümee ... 106
4.5.1 Schlussfolgerungen für die Erfassung des multidimensionalen Nutzerzustands ... 106
4.5.2 Erkenntnisse zu den Anforderungen aus dem Stand der Forschung ... 107

5 Experiment 2 – Retest zur Untersuchung der zeitlichen Stabiltät ... 109

5.1 Forschungsziele ... 109
5.1.1 Untersuchung der zeitlichen Stabilität der Eyetracking- und EEG-Maße ... 109
5.1.2 Untersuchung der diagnostischen Fähigkeiten des Zephyr BioHarness 3 ... 109
5.2 Methodisches Vorgehen ... 110
5.2.1 Versuchsdesign ... 110
5.2.2 Sensor Zephyr BioHarness 3 ... 111
5.2.3 Hypothesen ... 112
5.2.4 Datenaufbereitung und Auswertung ... 114
5.3 Ergebnisse zur zeitlichen Stabilität der physiologischen und verhaltensbasierten Maße 115
5.3.1 Test-Retest Korrelationen ... 115
5.3.2 Korrelation der Eyetracking- und EEG-Maße mit der Leistung und dem subjektiven Nutzerzustand in Test und Retest ... 116
5.4 Ergebnisse zu den Maßen des BioHarness ... 118
5.5 Diskussion ... 119
5.6 Resümee ... 121

6 Konzeption und Umsetzung der multifaktoriellen Echtzeitdiagnose RASMUS ... 123

6.1 Diagnosekonzept ... 123
6.1.1 Bestimmung von Adaptierungsbedarf ... 123
6.1.2 Bestimmung der Problemzustände und Ursachen ... 124
6.1.3 Überblick über das Diagnosekonzept ... 125
6.2 Beschreibung des Diagnoseprozesses ... 126
6.2.1 Daten erfassen und zusammenführen (Schritt 1) ... 127
6.2.2 Adaptierungsbedarf identifizieren (Schritt 2) ... 128
6.2.3 Kritische Zustände und Ursachen diagnostizieren (Schritt 3) ... 129
6.2.4 Diagnoseergebnisse bereitstellen (Schritt 4) ... 130
6.3 Experimentalparadigma für die Umsetzung ... 131
6.4 Betrachtete Nutzerzustände ... 133
6.5 Experiment 3 – Post hoc-Analyse zur Umsetzung einer Echtzeitdiagnose ... 134
6.5.1 Versuchsaufbau ... 134
6.5.2 Abhängige Variablen ... 135
6.5.3 Hypothesen und Forschungsfragen ... 136
6.5.4 Stichprobe und Durchführung ... 137
6.5.5 Datenaufbereitung und -auswertung ... 138
6.5.6 Ergebnisse ... 138
6.5.7 Diskussion ... 139
6.5.8 Untersuchung zum Verhalten der Diagnosemaße im zeitlichen Verlauf ... 140
6.6 Indikatoren und Regeln der Echtzeitdiagnose ... 141
6.6.1 Indikatoren und Regeln zur Diagnose von Leistungseinbrüchen ... 142

6.6.2 Indikatoren und Regeln zur Diagnose der Nutzerzustände 142
6.7 Technische Umsetzung 144

7 Experiment 4 – Validierung der multifaktoriellen Echtzeitdiagnose RASMUS 145
7.1 Methodische Umsetzung 145
7.1.1 Experimentalumgebung 145
7.1.2 Versuchsdesign 146
7.1.3 Versuchsaufbau 147
7.1.4 Vergleichsmaße 148
7.1.5 Hypothesen 149
7.1.6 Versuchsdurchführung 150
7.1.7 Stichprobe 151
7.1.8 Datenaufbereitung und -auswertung 151
7.2 Ergebnisse 152
7.2.1 Deskriptive Analyse 152
7.2.2 Hypothesenprüfende Untersuchung 155
7.2.3 Analyse zur Güte der diagnostischen Entscheidung 156
7.3 Bewertung der Ergebnisse 160
7.3.1 Diagnose kritischer Beanspruchung 160
7.3.2 Diagnose kritischer Müdigkeit 161
7.3.3 Diagnose kritischer Aufmerksamkeit 161
7.3.4 Leistungseinbrüche ohne Diagnoseergebnis 162
7.3.5 Einflussfaktoren auf die Ergebnisse des Validierungsexperiments 163
7.3.6 Resümee 163
7.4 Weiterführende Arbeiten zur Optimierung der Diagnose 164

8 Abschließende Diskussion der Forschungsarbeit und Ausblick 167
8.1 Bewertung des Diagnosekonzepts 167
8.1.1 Bezug der konzeptionellen Maßnahmen zu Zielsetzungen und Anforderungen 167
8.1.2 Erkenntnisse aus dem Validierungsexperiment 168
8.1.3 Vergleich mit dem Forschungsvorhaben „Human Performance Envelope“ 169
8.2 Ausblick 170
8.2.1 Erweiterung der Diagnosefähigkeiten 171
8.2.2 Einsatzmöglichkeiten der multifaktoriellen Echtzeitdiagnose 172

9 Literatur 173

Anhang A. Versuchsmaterialien und ergänzende Auswertungen zu Experiment 1 199
A.1 Versuchsplan (Experiment 1) 199
A.2 Einverständniserklärung (Experiment 1) 200
A.3 Instruktion (Experiment 1) 201
A.4 Versuchsprotokoll (Experiment 1) 204
A.5 Fragebogen SAM - Self-Assessment-Manikin (Experiment 1) 206
A.6 Fragebogen zu individuellen Faktoren und Personenangaben (Experiment 1) 207
A.7 Fragebogen zum Situationsbewusstsein und NASA-TLX (Experiment 1) 208
A.8 Abschlussfragen (mündlich) (Experiment 1) 210
A.9 Aufzeichnungsqualität der Eyetracking-und EEG-Daten (Experiment 1) 211

A.10 Detailauswertung NASA-TLX (Experiment 1) 212
A.11 Interaktionsdiagramme zur varianzanalytischen Auswertung in Abschnitt 4.3.1 (Experiment 1) 213

Anhang B. Instruktion zu Experiment 2 214

Anhang C. Versuchsmaterialien und ergänzende Auswertungen zu Experiment 3 215

C.1 Aufgabenbeschreibung und ID-Kriterien (Experiment 3) 215
C.2 Beschreibung der Konsolenbedienung (Experiment 3) 216
C.2 Ergebnisse auf individueller Ebene (Experiment 3) 219

Anhang D. Darstellungen zur technischen Umsetzung der Echtzeitdiagnose 220

D.1 Architektur der Soft- und Hardwarekomponenten 220
D.2 Regeleditor 221
D.3 Leistungs- und Zustandsmonitor 222

Anhang E. Versuchsmaterialien und ergänzende Auswertungen zu Experiment 4 223

E.1 Instruktion (Experiment 4) 223
E.2 Versuchsprotokoll (Experiment 4) 226
E.3 Zusatzauswertungen (Experiment 4) 227

Abbildungsverzeichnis

Abbildung 1. MMI-Modell mit aktiver Wechselwirkung zwischen Operateur und technischer Systemkomponente (Fuchs & Schwarz, 2014; Schwarz & Fuchs, 2014) 7

Abbildung 2. Arbeitsplan und Kapitelstruktur (Kapitelnummern in Klammern) 11

Abbildung 3. Einflussfaktoren auf die psychische Beanspruchung nach DIN EN 10 075-1 (2000) 41

Abbildung 4. Modell zur Multiple Resource Theory nach Wickens (1984) 42

Abbildung 5. Die drei Dimensionen des Cognitive Task Load-Modells von Neerincx (2003) basierend auf einer Darstellung in DeGreef et al. (2009) 43

Abbildung 6. Darstellung des Zusammenhangs zwischen Aufgabenanforderungen und der Leistung nach Veltman & Jansen (2006) 44

Abbildung 7. Circumplex-Modell basierend auf Russell, 1980 46

Abbildung 8. Vereinfachte Darstellung des Pfadmodells von Nicholls et al. (2012) 47

Abbildung 9. Modell von Porter & Lawler (1968) nach Pelz (2004) 50

Abbildung 10. Differenzierung zwischen schlafbezogener und aufgabenbezogener Müdigkeit nach May & Baldwin (2009) 52

Abbildung 11. SEEV-Modell zur Aufmerksamkeit basierend auf Wickens & McCarley (2008) 55

Abbildung 12. Modell zum Situationsbewusstsein nach Endsley (1995) 58

Abbildung 13. Darstellung des multidimensionalen Nutzerzustands 60

Abbildung 14. Wirkzusammenhänge zwischen individuellen Faktoren und dem Nutzerzustand basierend auf Veltman et al. (2004) 61

Abbildung 15. Generisches Modell zu Einflussfaktoren und Auswirkungen des Nutzerzustands 65

Abbildung 16. Modell mit den in Experiment 1 untersuchten Komponenten zur Nutzerzustandsbewertung 70

Abbildung 17. Verwendete Radarsimulation in Experiment 1 71

Abbildung 18. Versuchsdesign (Experiment 1) 74

Abbildung 19. Übersicht über die in Experiment 1 erfassten unabhängigen und abhängigen Variablen 75

Abbildung 20. Aufgabe aus dem „Links-Rechts-Test“ 77

Abbildung 21. Eyetracker Tobii X120 (links) und das EEG Headset Emotiv EPOC^{TM} (rechts) 79

Abbildung 22. Hypothetisierte Abhängigkeiten und Zusammenhänge zwischen den Variablen (Experiment 1) 83

Abbildung 23. Versuchsaufbau (Experiment 1) 87

Abbildung 24. Mittelwerte pro Faktorstufe für die subjektiven Bewertungen in Experiment 1 (Fehlerbalken: Standardfehler) 92

Abbildung 25. Mittelwerte pro Faktorstufe für die Eyetracking-Maße in Experiment 1 (Fehlerbalken: Standardfehler) 94

Abbildung 26. Mittelwerte pro Faktorstufe für die Leistungsmaße in Experiment 1 (Fehlerbalken: Standardfehler) 95

Abbildung 27. Individualkorrelationen pro Versuchsperson (VP) zwischen dem NASA-TLX-Gesamtscore und der Pupillenweite (a), der Fixationsdauer (b) und dem EEG-Engagement (c) sowie zwischen der Subskala Frustration und EEG-Frustration (d) (** p<.01; * p<.05 bei einseitigem Testen) – Experiment 1 98

Abbildung 28. Individualkorrelationen pro Versuchsperson (VP) zwischen den Eyetracking- und EEG-Maßen und dem Punktestand (** *p*<.01; * *p*<.05 bei einseitigem Testen) – Experiment 1 99

Abbildung 29. Individualkorrelationen pro Versuchsperson (VP) zwischen dem Punktestand und dem NASA-TLX-Gesamtscore (** *p*<.01, **p*<.05) 101

Abbildung 30. Experimentelles Design für die Retest-Untersuchung (Experiment 2) 111

Abbildung 31. Multisensor Zephyr BioHarness 3 mit einer Übersicht über die wesentlichen von dem Sensor erfassten physiologischen und verhaltensbasierten Maße (Originalbezeichnung in Klammern) 112

Abbildung 32. Test-Retest-Korrelationen für die Eyetracking- und EEG-Maße pro Teilnehmer (Ø = Mittelwert der Individualkorrelationen) – Experiment 2 116

Abbildung 33. Individualkorrelationen und Mittelwert (Ø) der Eyetracking- und EEG-Maße mit der Leistung (Punktestand) für das erste Experiment und den Retest (Experiment 2) 117

Abbildung 34. Mittelwerte und Standardfehler pro Faktorstufe für die Maße des BioHarness in Experiment 2 (Fehlerbalken: Standardfehler; *p<.05; ns – nicht signifikant) 119

Abbildung 35. Vorgehen bei der multifaktoriellen Nutzerzustandsdiagnose in RASMUS 125

Abbildung 36. Schritte der Echtzeitdiagnose als Bestandteil der Zustandsregulierung des adaptiven Systems 126

Abbildung 37. Aktivitätsdiagramm zur Erfassung und Zusammenführung der Daten (Schritt 1) 127

Abbildung 38. Aktivitätsdiagramm zur Bestimmung des Adaptierungsbedarfs (Schritt 2) 128

Abbildung 39. Aktivitätsdiagramm zur Ursachenanalyse (Schritt 3) 130

Abbildung 40. Aktivitätsdiagramm zur Bereitstellung der Diagnoseergebnisse (Schritt 4) 131

Abbildung 41. GUI zur Bearbeitung der Experimentalaufgabe 132

Abbildung 42. Mittelwerte und Standardfehler des NASA-TLX-Gesamtscore, der Fixationsdauer, der Pupillenweite und der Anzahl Mausklicks pro Versuchsbedingung (Experiment 3) 139

Abbildung 43. Zeitlicher Verlauf der Variablen Pupillenweite, Anzahl Mausklicks und Anzahl Aufgaben pro Szenario gemittelt über 30-Sekunden-Intervalle (Experiment 3) 141

Abbildung 44. a: Versuchsaufbau mit Probandenarbeitsplatz; b: Arbeitsplätze des Versuchsleiters und des Probanden (Experiment 4) 145

Abbildung 45. Phasen im Versuch (Experiment 4) 147

Abbildung 46. Darstellung der in Experiment 4 verwendeten KU-Skala (Heller, 1982) 148

Abbildung 47. Durchschnittliche Zeitanteile hoher Beanspruchung und passiver aufgabenbezogener Müdigkeit pro Phase in Prozent (Fehlerbalken: Standardfehler) – Experiment 4 153

Abbildung 48. Verteilung von Leistungseinbrüchen im Szenarioverlauf (Experiment 4) 153

Abbildung 49. Durchschnittliche Abweichungen der subjektiven Bewertungen von der Baseline für kritische und unkritische Ausprägungen der Nutzerzustände nach RASMUS (Fehlerbalken: Standardfehler) – Experiment 4 155

Abbildung 50. Prozentuale Häufigkeit von nicht korrektem Level-1-SA bei Leistungseinbrüchen (LE) mit kritisch und nicht kritisch bewerteter Aufmerksamkeit durch RASMUS (Experiment 4) 156

Abbildung 51. Anteil gültiger Messwerte pro Aufzeichnung für die Eyetracking-Daten (links) und die Daten des Klassifikators Emotiv-Engagement (rechts) in absteigender Reihenfolge sortiert 211

Abbildung 52. Unterschiede zwischen den Faktorstufen auf den Subskalen des NASA-TLX 212

Abbildung 53. Interaktionsdiagramme zur Interaktion zwischen Kooperativität und Lärm hinsichtlich des NASA-TLX-Gesamtscore 213

Abbildung 54. Interaktionsdiagramme zur Interaktion zwischen Anzahl Areas und Kooperativität hinsichtlich der Variable Ausgelassene Rechenaufgaben 213

Abbildung 55. Mittelwerte der betrachteten Nutzerzustandsindikatoren in den Versuchsbedingungen Underload, Normal Load und Overload pro Versuchsperson (VP) 219

Abbildung 56. Häufigkeit der Leistungseinbrüche (LE) mit (un-)kritischer Beanspruchung, (un-)kritischer Müdigkeit und (un-)kritischer Aufmerksamkeit pro Person 227

Abbildung 57. Durchschnittliche Abweichungen der subjektiven Bewertung von der Baseline bei (un-)kritischer Beanspruchung, (un-)kritischer Müdigkeit und (un-)kritischer Aufmerksamkeit pro Person 228

Abbildung 58. Baseline-Abweichungen der subjektiven Bewertungen bezüglich Anstrengung, Müdigkeit und Aufmerksamkeit bei Leistungseinbrüchen ohne kritisch diagnostizierten Nutzerzustand 229

Abbildung 59. Baseline-Abweichungen der subjektiven Bewertungen bezüglich Motivation und den Dimensionen des SAM bei Leistungseinbrüchen ohne kritisch diagnostizierten Nutzerzustand 229

Abbildung 60. Anzahl kritischer Indikatoren bei Leistungseinbrüchen mit kritischer Beanspruchung und kritischer Müdigkeit 230

Abbildung 61. Prozentanteile kritischer Indikatorausprägungen bei Leistungseinbrüchen mit kritischer Beanspruchung und kritischer Müdigkeit 230

Tabellenverzeichnis

Tabelle 1. Übersicht über die zu untersuchenden Forschungsfragen 10

Tabelle 2. Übersicht über betrachtete Nutzerzustände, Anwendungsbereiche und verwendete Methoden in Studien zu adaptiven Systemen 16

Tabelle 3. Übersicht über Vor- und Nachteile der Gruppen unterschiedlicher Erfassungsmethoden 26

Tabelle 4. Übersicht über okulomotorische Maße und ihren Zusammenhang mit verschiedenen Nutzerzuständen 27

Tabelle 5. Übersicht über Maße der Hirnaktivität und ihren Zusammenhang mit verschiedenen Nutzerzuständen 29

Tabelle 6. Übersicht über peripherphysiologische und verhaltensbasierte Maße und ihren Zusammenhang mit verschiedenen Nutzerzuständen (in alphabetischer Reihenfolge) .. 30

Tabelle 7. Extrahierte Anforderungen an die Konzeption einer Echtzeitdiagnose 33

Tabelle 8. Übersicht über die sechs in der Dissertation betrachteten Nutzerzustände und ihre Auswirkungen auf die Leistung 39

Tabelle 9. Erkenntnisse aus den Theorien und Modellen zur mentalen Beanspruchung 45

Tabelle 10. Erkenntnisse aus den Theorien und Modellen zum emotionalen Zustand 48

Tabelle 11. Erkenntnisse aus den Theorien und Modellen zur Motivation 51

Tabelle 12. Erkenntnisse aus den Theorien und Modellen zur Müdigkeit 53

Tabelle 13. Erkenntnisse aus den Theorien und Modellen zur Aufmerksamkeit und Vigilanz 57

Tabelle 14. Erkenntnisse aus den Theorien und Modellen zum Situationsbewusstsein 59

Tabelle 15. Merkmale der Realaufgabe und Repräsentation durch die Experimentalaufgabe 71

Tabelle 16. Testabfolge pro Versuchsperson in Experiment 1 (Ausschnitt) 75

Tabelle 17. Beschreibung der Metriken von Emotiv EPOC^{TM} 81

Tabelle 18. Angenommene Mittelwertunterschiede zwischen den Faktorstufen für die subjektiven Maße (H1a), die physiologischen und verhaltensbasierten Maße (H1b) und die Leistungsmaße (H1c) – Experiment 1 84

Tabelle 19. Angenommene Richtung der Korrelationen für die in den Hypothesen H2a und H2b erwarteten (Kausal-)Zusammenhänge (Experiment 1) 85

Tabelle 20. Angenommene Richtung der Korrelationen für die in H3 erwarteten Zusammenhänge zwischen den Diagnosemaßen (Experiment 1) 86

Tabelle 21. Versuchsabfolge pro Versuchssitzung (Experiment 1) 87

Tabelle 22. Ergebnisse der varianzanalytischen Auswertung bezüglich der Fragebögen NASA-TLX und SAM (Experiment 1) 92

Tabelle 23. Ergebnisse der varianzanalytischen Auswertung bezüglich Eyetracking- und EEG-Maße in Experiment 1 94

Tabelle 24. Ergebnisse der varianzanalytischen Auswertung bezüglich der Leistungsmaße (Experiment 1) 95

Tabelle 25. Spearman-Rho-Korrelationskoeffizienten für die Korrelationen zwischen individuellen Faktoren und Indikatoren des Nutzerzustands (Experiment 1) 96

Tabelle 26. Produkt-Moment-Korrelationen zwischen subjektiven Maßen und den Eyetracking-/ EEG-Maßen auf Gruppenebene unkorrigiert (r) und nach Bereinigung von Zwischensubjekteffekten (r_{part}) – Experiment 1 97

Tabelle 27. Produkt-Moment-Korrelationen zwischen dem Punktestand und den Eyetracking- / EEG-Maßen auf Gruppenebene unkorrigiert (*r*) und nach Bereinigung von Zwischensubjekteffekten (r_{part}) – Experiment 1 99

Tabelle 28. Ergebnisse der regressionsanalytischen Untersuchungen zur Vorhersage der Leistung durch einzelne und kombinierte Eyetracking- und EEG-Maße (Experiment 1) ... 100

Tabelle 29. Produkt-Moment-Korrelationen zwischen dem Punktestand und den subjektiven Maßen (NASA-TLX und SAM) auf Gruppenebene unkorrigiert (r) und nach Bereinigung von Zwischensubjekteffekten (r_{part}) – Experiment 1 101

Tabelle 30. Ergebnisübersicht zum Einfluss der Anforderungsmerkmale auf den diagnostizierten Nutzerzustand (Hypothese 1 – Experiment 1) 102

Tabelle 31. Ergebnisübersicht zu den Korrelationen zwischen individuellen Faktoren und den Nutzerzustands- und Leistungsmaßen (Hypothese 2 – Experiment 1) 104

Tabelle 32. Ergebnisübersicht zu den Korrelationen zwischen den Diagnosemaßen (Hypothese 3 – Experiment 1) 105

Tabelle 33. Erkenntnisse zu den in Experiment 1 berücksichtigten Anforderungen aus Abschnitt 2.5 108

Tabelle 34. Hypothesen zur zeitlichen Stabilität für die Eyetracking- und EEG-Maße (Experiment 2) 112

Tabelle 35. Hypothetisierte Zusammenhänge für die Maße des BioHarness (Experiment 2) 114

Tabelle 36. Test-Retest Korrelationen der Eyetracking- und EEG-Maße auf Gruppenebene (Experiment 2) 116

Tabelle 37. Statistische Kennwerte zu den Korrelationen der Eyetracking- und EEG-Maße mit der Leistung (Punktestand) und dem NASA-TLX auf Gruppenebene im ersten Experiment und im Retest (Experiment 2) 117

Tabelle 38. Korrelationen der Maße des BioHarness mit dem NASA-TLX und dem Punktestand (Leistung) auf Gruppenebene nach Bereinigung von Zwischensubjekteffekten (Experiment 2) 119

Tabelle 39. Beschreibung der in der Experimentalaufgabe verwendeten Teilaufgaben und Angabe ihrer Prioritäten 133

Tabelle 40. Anzahl der Teilaufgaben pro Versuchsszenario (Belastungsstufe) – Experiment 3 135

Tabelle 41. Versuchsdesign (Experiment 3) 135

Tabelle 42. Angenommene Mittelwertunterschiede zwischen den Faktorstufen Underload (UL), Normal Load (NL) und Overload (OL) für die betrachteten Indikatoren in Experiment 3 136

Tabelle 43. Ergebnisse zu Haupteffekten und Paarvergleichen zwischen den Belastungsstufen (Experiment 3) 138

Tabelle 44. Regeln zur Bestimmung von Leistungseinbrüchen (LE) pro Teilaufgabe 142

Tabelle 45. Indikatoren und Regeln zur Diagnose kritischer Nutzerzustände 142

Tabelle 46. Darstellung des quasiexperimentellen Versuchsdesigns mit den möglichen Diagnoseergebnissen von RASMUS bei Leistungseinbrüchen (LE) und der subjektiven Bewertung des Nutzerzustands als Vergleichsmaß 146

Tabelle 47. Subjektive Maße zur Erfassung der Nutzerzustandsdimensionen (Experiment 4) 148

Tabelle 48. Hypothesen zur Validierung der von RASMUS diagnostizierten Nutzerzustände bei Leistungseinbrüchen (Experiment 4) 150

Tabelle 49. Häufigkeiten diagnostizierter kritischer und unkritischer Nutzerzustände bei Leistungseinbrüchen in Experiment 4 154

Tabelle 50. Ergebnisse der inferenzstatistischen Auswertung zur Überprüfung der Hypothesen H1-H3a (Experiment 4) 156

Tabelle 51. Diagnostische Kennwerte zur Güte der Diagnosefähigkeiten von RASMUS (bei Verwendung der dichotomisierten subjektiven Vergleichsmaße als Referenz) – Experiment 4 158

Tabelle 52. Ergebnisse der ROC-Kurvenanalyse bei Verwendung der Baseline und dem bestmöglichen Trennwert zur Diskriminierung zwischen kritischen und unkritischen Nutzerzuständen (Experiment 4) 159

Tabelle 53. Übersicht über die Ergebnisse des Validierungsexperiments (Experiment 4) 163

Tabelle 54. Berücksichtigung der Anforderungen aus der Literaturanalyse (Kapitel 2) im Diagnosekonzept von RASMUS 168

Tabelle 55. Ergebnisse der Varianzanalyse für den Gesamtscore und die Subskalen des NASA-TLX 212

Glossar

AAW	Anti Air Warfare: Verteidigung des eigenen Luftraums gegen das Eindringen feindlicher Flugzeuge und anderer Flugkörper.
Adaptierung	Anpassung des Verhaltens an Umwelt- und Nutzerzustandsveränderungen
ADAM	Advanced Dynamic Adaptation Management: Komponente innerhalb des zu entwickelnden adaptiven Systems, welche die Auswahl, Konfiguration und Anwendung von Adaptierungsstrategien vornimmt.
AOI	Area of Interest: Bereiche auf einer Benutzungsoberfläche, für die separate Analysen der Eyetrackingdaten vorgenommen werden, z.B. um Dauer und Häufigkeit der Betrachtung zu bestimmen.
API	Application Programming Interface; Programmierschnittstelle eines Systems zur Anbindung von systemunabhängigen Programmen
Arousal	unspezifischer physiologischer Erregungszustand, der u.a. zur Beschreibung emotionaler Zustände verwendet wird
AUC	Gebiet unterhalb einer ROC-Kurve, die zur Bewertung der diagnostischen Qualität eines Diagnoseinstruments herangezogen werden kann
EEG	Elektroenzephalogramm – Messung der Hirnaktivität
EKG	Elektrokardiogramm – Messung der Herztätigkeit
ERP	Event-related potentials – ereigniskorrelierte Potenziale, die aus dem EEG abgeleitet werden können und in Zusammenhang mit einem beobachtbaren Reiz stehen
fNIRS	Funktionelle Nahinfrarot-Spektroskopie; bildgebendes Verfahren zur Aufzeichnung der Hirnaktivität
GUI	grafische Benutzungsoberfläche, die eine Visualisierung von Informationen und eine Interaktion mit dem technischen System ermöglicht.
ICC	Intraclass correlation; Koeffizient, der die Übereinstimmung von zwei oder mehr Beobachtungen an den gleichen Fällen anzeigt und zur Abschätzung der Reliabilität von Messwertreihen verwendet werden kann
ISR	Identification Safety Range: Sicherheitszone im Umkreis eines Marineschiffs
I-VT Filter	Velocity-Threshold Identification; Klassifikationsalgorithmus für Fixationen, der Klassifizierungen anhand der Geschwindigkeit der Blickbewegungen vornimmt.
Mentale Beanspruchung	Eine Form der psychischen, nicht körperlichen Beanspruchung, die durch aufgabenspezifische Belastungsfaktoren wie die Schwierigkeit und Komplexität der Aufgabe hervorgerufen wird.
MMI	Mensch-Maschine-Interaktion: aufeinander bezogenes Handeln von Mensch und Technik
Negativer prädiktiver Wert	Errechnet sich aus: richtig Negative / (richtig Negative + falsch Negative) und bezeichnet den Anteil an richtigen Klassifizierungen unter allen negativen Befunden.
NRTT	Non Real-Time Track: Kontakt, der manuell auf dem Lagebild angelegt wird

Positiver prädiktiver Wert	Errechnet sich aus: richtig Positive / (richtig Positive + falsch Positive) und bezeichnet den Anteil an richtigen Klassifizierungen unter allen positiven Befunden.
Prävalenz	Auftretenshäufigkeit eines Zustands in der betrachteten Grundgesamtheit
RASMUS	Real-Time Assessment of Multidimensional User State: Komponente innerhalb des zu entwickelnden adaptiven Systems, die Informationen über Leistungseinbrüche, kritische Nutzerzustände und Einflussfaktoren bereit stellt
Regeleditor	Softwaretool, in dem die in RASMUS und ADAM verwendeten Regeln zur Nutzerzustandsdiagnose und Adaptierung erstellt, konfiguriert und für die Anwendung aktiviert werden können.
ROC-Kurven	Grafisches Beurteilungsverfahren der Güte eines Klassifikators. Dabei wird die Sensitivität für verschiedene Klassifikatorwerte (bei einem dichotomen Klassifikator ist dies nur ein Wert) auf der Y-Achse und die falsch positive Rate (1-Spezifität) auf der X-Achse eines Koordinatensystems abgetragen. Durch Verbindung der Punkte mit Anfang und Ende des Koordinatensystems entsteht die in der Regel nach oben gewölbte Receiver Operating Characteristic (ROC)-Kurve. Eine ROC-Kurve nahe der Diagonalen weist darauf hin, dass die Güte des Klassifikators weitgehend der Güte zufälliger Klassifizierungen entspricht.
SA	Situationsbewusstsein (engl. Situation Awareness); Endsley (1995) unterscheidet drei Ebenen von SA: Level 1 - Wahrnehmen, Level 2 - Verstehen und Level 3 - Projektion in die Zukunft
SAGAT	Situation Awareness Global Assessment Technique: Verfahren von Endsley (1988) zur Erfassung des Situationsbewusstseins in Simulationsumgebungen.
SDK	Software Development Kit: enthält Informationen, Dokumentationen sowie Werkzeuge für die Software-Entwicklung
SDNN	Standardabweichung des Interbeat-Intervalls; SDNN stellt ein zeitbezogenes Maß zur Bestimmung der Herzratenvariabilität dar.
Sensitivität	Errechnet sich aus: richtig Positive / (richtig Positive + falsch Negative) und bezeichnet den Anteil an positiven Fällen in der Grundgesamtheit, die richtig als positiv klassifiziert wurden.
Spezifität	Errechnet sich aus: richtig Negative / (richtig Negative + falsch Positive) und bezeichnet den Anteil an negativen Fällen in der Grundgesamtheit, die richtig als negativ klassifiziert wurden.
TDA	Taktisches Lagebild, das die durch Radar erfassten Objekte in der Umgebung anzeigt.
Valenz	Dimension zur Bewertung emotionaler Zustände in Hinblick auf das mit ihnen verbundene positive oder negative Erleben
Vigilanz	Aufrechthaltung der Aufmerksamkeit über einen ausgedehnten Zeitraum bei in der Regel monotonen Überwachungstätigkeiten
WR	Waffenreichweite z.B. eines Marineschiffs

1 Einführung

„Errare humanum est“ – „Irren ist menschlich“ besagt eine bekannte lateinische Redewendung. Tatsächlich sind mehrheitlich menschliche Fehler beteiligt, wenn es in Mensch-Maschine-Systemen zu Unglücken oder unerwünschten Zwischenfällen kommt. Im Luftverkehr wird der Anteil an Flugzeugunfällen, die mit menschlichem Fehlverhalten verbunden sind, auf 70-80 Prozent geschätzt (Wiegmann & Shappell, 2003). In den USA sind einer Studie zufolge über 90 Prozent der Autounfälle auf menschliche Fehler zurückzuführen (Singh, 2015). Für Zwischenfälle in Kernkraftwerken wurde ein Anteil von 65 Prozent ermittelt (Dhillon, 2014 bezugnehmend auf Trager, 1985).

Um die Sicherheit in Mensch-Maschine-Systemen zu erhöhen, erscheint es somit ratsam, die Ursachen für menschliches Fehlverhalten zu identifizieren und diesen durch geeignete technische Maßnahmen entgegenzuwirken. Neben anderen Einflussfaktoren nimmt der mentale Zustand des Nutzers (Nutzerzustand) unmittelbar Einfluss auf die menschliche Leistungsfähigkeit und kann bei ungünstiger Ausprägung Leistungsdefizite und Fehlverhalten provozieren (Wiegmann & Shappell, 2001). Das nachfolgend beschriebene Unglück von Air-France-Flug AF447 ist ein eindrucksvolles Beispiel dafür, welche Ursachen und Auswirkungen kritische Nutzerzustände in sicherheitsrelevanten Mensch-Maschine-Systemen haben können.

1.1 Flug AF447

Am 1. Juni 2009 stürzte die Maschine von Air-France-Flug AF447 auf dem Weg von Rio de Janeiro nach Paris in den Atlantischen Ozean. Alle 228 Passagiere und Crewmitglieder kamen dabei ums Leben. Zwei Jahre nach dem Unglück wurden die Flugschreiber gefunden. Sie ermöglichten es, den Unfallhergang detailliert zu rekonstruieren und in einem Flugunfallbericht zu dokumentieren (BEA, 2012).

1.1.1 Einflussfaktoren auf den Nutzerzustand

Im Flugunfallbericht werden verschiedene Faktoren genannt, die sich negativ auf den mentalen Zustand der Piloten auswirkten und vermutlich entscheidend dazu beigetragen haben, dass die Piloten die Kontrolle über das Flugzeug verloren (BEA, 2012):

- *Technischer Defekt:* Die Pitot-Sonden, die die Geschwindigkeit des Flugzeugs messen, waren vereist und konnten dadurch keine verlässliche Geschwindigkeitsanzeige mehr liefern. Dies war auch mit einem Verlust der Höheninformation verbunden. Den Piloten war der Grund für diese Anomalien jedoch nicht bekannt.
- *Abschalten des Autopiloten:* Die Geschwindigkeitsanomalie führte dazu, dass sich der Autopilot abschaltete, so dass die Piloten selbst die Steuerung des Flugzeugs übernehmen mussten. Der Grund für das Abschalten des Autopiloten war den Piloten ebenfalls nicht bekannt.

- Die *manuelle Steuerung* des Flugzeugs in Reiseflughöhe war für die Piloten eine ungewohnte Situation. Diese Aufgabe wird während des Flugs üblicherweise vom Autopiloten übernommen und daher selten trainiert.
- Aufgrund *schlechter Sichtverhältnisse* (Nacht und Gewitter) konnten sich die Piloten nicht am Horizont orientieren, um die Fluglage zu bestimmen.
- *Turbulenzen* erschwerten das manuelle Fliegen sowie die Orientierung.
- Durch das Abschalten des Autopiloten und die Geschwindigkeitsanomalien wurden eine Reihe von *Alarmen* ausgelöst, die sich nicht von alleine wieder abstellten.
- „Pilot Flying" (PF) war der Pilot mit der *geringsten Flugerfahrung*. Der Kapitän machte zu der Zeit, als sich der Autopilot abschaltete, eine Ruhepause und kam erst 2,5 Minuten vor dem Absturz in das Cockpit zurück.

Mit Hilfe der Aufnahmen des Stimmenrekorders konnte rekonstruiert werden, dass diese Faktoren den mentalen Zustand der Piloten in verschiedener Hinsicht negativ beeinflussten und damit ein Fehlverhalten der Piloten begünstigten.

1.1.2 Auswirkungen auf den emotionalen Zustand

Dem Flugunfallbericht zufolge löste das plötzliche Abschalten des Autopiloten bei den Piloten einen emotionalen Schock aus. Wohl aufgrund dieses Überraschungseffekts und der geringen Erfahrung im Fliegen in großer Höhe zog der PF die Maschine stark nach oben und brachte sie damit in den Steigflug: *"The excessive nature of the PF's inputs can be explained by the startle effect and the **emotional shock** at the autopilot disconnection, amplified by the lack of practical training for crews in flight at high altitude"* (BEA, 2012, S. 173).

1.1.3 Auswirkungen auf die Beanspruchung

Der Steigflug hatte zur Folge, dass die Maschine immer mehr an Geschwindigkeit verlor, was letztendlich zu einem Strömungsabriss führte. Es wird jedoch vermutet, dass die Piloten im Cockpit das geänderte Verhalten der Maschine durch die schlechten Sichtbedingungen und die Turbulenzen nicht richtig wahrnehmen konnten. Zudem waren die Piloten damit beschäftigt, die Ursache für das Abschalten des Autopiloten herauszufinden. Insbesondere der PF schien damit überfordert zu sein, die Maschine zu steuern und gleichzeitig ein Verständnis für die Situation zu entwickeln: *"The PF may therefore have been **overloaded** by the combination of his immediate and natural attempts to understand the situation that was added to the already demanding task of handling the aeroplane."* (BEA, 2012, S. 176).

1.1.4 Auswirkungen auf das Situationsbewusstsein

Eine weitere Schwierigkeit, das Flugzeug unter Kontrolle zu bekommen und zu stabilisieren, bestand darin, dass den Piloten durch das Vereisen der Pitot-Sonden keine verlässlichen Geschwindigkeitsinformationen zur Verfügung standen. Es wird im Flugunfallbericht vermutet, dass der PF dadurch ein falsches mentales Modell von der Flugsituation entwickelt hatte und

dachte, die Geschwindigkeit des Flugzeugs sei zu hoch: *"In the absence of airspeed information known to be reliable, it is possible that the PF* ***thought that the aeroplane was in an overspeed situation****"* (BEA, 2012, S. 179). Dies könnte erklären, warum der PF das Flugzeug durch „nose-up inputs" immer weiter nach oben steuerte und sich die Geschwindigkeit dadurch immer weiter bis zum Strömungsabriss verringerte.

1.1.5 Auswirkungen auf die Aufmerksamkeit

Hinsichtlich der Alarme wird im Bericht bemerkt, dass mehrere Alarme ertönten, wobei ein Alarm („C-chord warning") sehr dominant war, da er mehr als 30 Sekunden lang dauerhaft ertönte. Das „STALL warning", das auf einen drohenden Strömungsabriss hinweist, ertönte hingegen nur kurz. Die Piloten schenkten dem STALL warning möglicherweise deshalb weniger Aufmerksamkeit. Auch auf das „Buffet", einem Vibrieren das kurz vor einem Strömungsabriss auftritt, reagierten die Piloten nicht: *"The crew never referred either to the stall warning or the buffet that they likely felt. […] In an aural environment that was already saturated by the C-chord warning, the possibility that the crew* ***did not identify the stall warning*** *cannot be ruled out"* (BEA, 2012, S. 179).

1.1.6 Resümee

Das Beispiel dieses Flugzeugunglücks zeigt, dass es selbst bei modernen, hoch automatisierten Mensch-Maschine-Systemen wie einem Flugzeug zu Fehlverhalten bei Mensch und Technik kommen kann, das im ungünstigsten Fall Menschenleben fordert. Der Unfallhergang legt dabei nahe, dass Ursachen nicht allein bei der Performanz der Technik und der Operateure zu suchen sind. Vielmehr kommt auch der Gestaltung des technischen Systems und der Mensch-Maschine-Interaktion (MMI) eine entscheidende Bedeutung zu. So wurden die aufgeführten mentalen Problemzustände der Piloten erst durch die Interaktion mit dem technischen System hervorgerufen (plötzlicher Ausfall des Autopiloten, unzureichende Rückmeldung über Ursachen für den Ausfall, Vielzahl an Alarmen, keine Unterstützung bei manueller Steuerung). Sie hätten möglicherweise durch ein anderes Verhalten des technischen Systems vermieden oder reduziert werden können. Forschungsarbeiten, die sich mit Problemen bei hochautomatisierten Systemen und deren Lösungsmöglichkeiten beschäftigen (siehe nächsten Abschnitt), bestätigen diese These.

1.2 Probleme bei hochautomatisierten Systemen

Dass die zunehmende Automatisierung technischer Systeme nicht nur positive Effekte auf die Leistungsfähigkeit des Menschen mit sich bringt, wurde bereits vor einigen Jahrzehnten erkannt. Bainbridge (1983) identifizierte einige sogenannte „ironies of automation", also Widersprüchlichkeiten der Automatisierung. Eine Ironie besteht darin, dass Automation den Menschen zwar einerseits entlasten, unterstützen und teilweise auch ersetzen soll, dass die Automation jedoch auch fehlerhaft sein oder versagen kann. Dies führt dazu, dass dem Menschen bei Übernahme der Aufgaben durch die Automation dennoch eine Überwachungsrolle zukommt. Ironischerweise stellt die Ausführung dieser Überwachungsaufgabe ebenfalls eine Beanspruchung dar und die Aufgaben, die trotz Automation beim Menschen verbleiben, sind entsprechend schwieriger. Verschiedene Faktoren wirken sich bei dieser Aufgabenteilung negativ auf die Leistungsfähigkeit des

menschlichen Operateurs aus. Manzey (2008) fasst diese in drei Problemfeldern zusammen: Schwierigkeiten bei der Aufrechthaltung des Situationsbewusstseins, mangelndes oder übersteigertes Vertrauen in die Automation, sowie den Verlust manueller Fertigkeiten. Diese Aspekte werden im Folgenden näher ausgeführt.

1.2.1 Aufrechterhaltung des Situationsbewusstseins

Bei hochautomatisierten Systemen besteht eine wesentliche Herausforderung für den Menschen darin, dass er trotz seiner passiven Rolle aufmerksam bleiben und sein Situationsbewusstsein aufrecht erhalten muss, um auf etwaige Automationsfehler rechtzeitig und angemessen reagieren zu können. Dass dies nicht immer gelingt, verdeutlicht der Absturz von Flug AF447. So waren die Piloten nach dem Ausfall des Autopiloten nicht in der Lage, ein korrektes mentales Modell von der Situation zu entwickeln und das Flugzeug unter Kontrolle zu bringen.

Die Beobachtung, dass Operateure automatisierter Systeme Schwierigkeiten haben, bei Automationsversagen das Vorgehen des Systems nachzuvollziehen und die manuelle Kontrolle zu übernehmen, wird auch als „Out-of-the-Loop-Performance“(OOLP)-Problem (e.g. Endsley & Kiris, 1995, Sarter, 1991) bezeichnet. Die Bildung eines für die Übernahme dieser Aufgaben nötigen Situationsbewusstseins wird in automatisierten Systemen durch verschiedene Faktoren erschwert:

1. Dadurch, dass der Mensch die Aufgaben nicht mehr selbst durchführt, sondern die Rolle eines passiven Überwachers einnimmt, ist er nicht mehr aktiv in den Regelkreis der Systemsteuerung eingebunden. Rückmeldekanäle, die bei manueller Ausführung für die Aufrechterhaltung des Situationsbewusstseins von Bedeutung sind, verändern sich dadurch oder fallen komplett weg (Manzey, 2008).
2. Es gilt zudem als erwiesen, dass sich mit zunehmender Dauer einer Überwachungsaufgabe die Vigilanzleistung drastisch verringert (vgl. Mackworth, 1948). Dies führt dazu, dass Informationen des technischen Systems, die für die Aufrechterhaltung des Situationsbewusstseins genutzt werden können, nach einer gewissen Zeit nicht mehr mit der nötigen Aufmerksamkeit wahrgenommen werden. Entsprechend haben Endsley & Kaber (1999) festgestellt, dass das Level-1-Situationsbewusstsein (Level-1-SA), das sich auf die Wahrnehmung von Informationen in der Umgebung bezieht, bei niedrigen Automationsgraden höher ist als bei hohen Automationgraden.
3. Hinzu kommt, dass automatisierte Systeme eine zusätzliche Komplexitätsschicht (Datenverarbeitung, Datenfusion und intelligente Steuerung) zwischen den tatsächlichen Systemprozessen und den durch den Nutzer wahrgenommenen und überwachten Daten erzeugen (Coury and Semmel, 1996). Es ist häufig nicht transparent für den Operateur, was das System gerade macht, so dass der Operateur kein korrektes mentales Modell über die Funktionsweise der Automation aufbauen kann. Dies kann in der Folge dazu führen, dass wahrgenommene Informationen nicht richtig interpretiert werden und falsche Schlussfolgerungen über künftige Prozesse abgeleitet werden. In der Literatur wird dies auch als „mode unawareness“ bezeichnet und gilt als häufiges Problem, das zu Flugzeugabstürzen beiträgt (Sarter & Woods, 1995).

1.2.2 Vertrauen in die Automation

Das Vertrauen in die Automation bestimmt wesentlich, inwieweit eine Automation in geeigneter Weise genutzt wird, so dass die Ziele der Automation erreicht werden können (Manzey, 2008). Angemessenes Vertrauen (im Englischen als „reliance“ oder „appropriate trust“ bezeichnet - Lee & See, 2004) liegt dann vor, wenn sich das Vertrauen des Operateurs in die Automation mit deren tatsächlichen Fähigkeiten deckt. Dies wird als wichtige Voraussetzung für ein adäquates Nutzungs- und Überwachungsverhalten des Menschen gesehen.

Probleme entstehen jedoch, wenn der Operateur ein übersteigertes Vertrauen (Overreliance) oder ein mangelndes Vertrauen (Underreliance) in die Automation aufweist (Parasuraman & Riley, 1997). Mangelndes Vertrauen führt Parasuraman & Riley (1997) zufolge zu einer zu geringen Nutzung der Automation. Die Folgen können darin bestehen, dass Warnungen und Alarme nicht ernst genommen werden und die angestrebte Beanspruchungsreduktion des Operateurs nicht erfolgt. Probleme mit mangelndem Vertrauen treten insbesondere bei nicht verlässlicher Automation auf. So kann eine hohe Anzahl von Fehlalarmen dazu führen, dass auch korrekte Alarme schließlich ignoriert werden (sog. „cry wolf-syndrome“, Wickens et al., 2016). In Hinblick auf das Air-France-Unglück wird von der BEA auch die Möglichkeit in Erwägung gezogen, dass die STALL-Warnung vor dem Absturz von den Piloten zwar gehört aber nicht ernst genommen wurde, da sie manchmal auch fälschlicherweise auslöst (BEA, 2012).

Demgegenüber kann eine hohe Zuverlässigkeit der Automation zu einem übersteigerten Vertrauen in die Automation beitragen und damit zu einer weiteren Ironie der Automation führen: Eine hohe Zuverlässigkeit verbessert zwar auf der einen Seite die Leistungsfähigkeit des Mensch-Maschine-Systems, auf der anderen Seite ist ein Eingreifen des Menschen bei hoher Zuverlässigkeit der Automation nur sehr selten erforderlich, so dass ein nachlässiges Überwachungs- und Kontrollverhalten meist keine negativen Konsequenzen nach sich zieht. Im Sinne einer positiven Rückkopplung kann dies die Tendenz zu einem übersteigerten Vertrauen in die Automation weiter verstärken (Manzey, 2008). Mögliche negative Konsequenzen – auch als „Complacency-Effekte“ bekannt – bestehen in einer unzureichenden Überwachung, einem Verlust des Situationsbewusstseins und einem Übersehen von Fehlern (Manzey, 2008, Manzey & Bahner, 2005).

1.2.3 Fertigkeitsverlust

Die dauerhafte Übernahme von Aufgaben durch das technische System (z.B. das Steuern und Stabilisieren des Flugzeugs während des Flugs durch den Autopiloten) bringt es mit sich, dass diese Aufgaben nur im Notfall durch den menschlichen Operateur ausgeführt werden müssen. Bainbridge (1983) weist in diesem Zusammenhang auf das Problem hin, dass die manuellen Fertigkeiten schlechter werden oder verloren gehen, wenn sie nicht benötigt werden. Die Folge ist, dass der Mensch in dem seltenen Fall, in dem ein manuelles Eingreifen notwendig ist, nicht mehr so performant reagieren kann. Bei Flug AF447 mussten die Piloten durch den Ausfall des Autopiloten das Flugzeug in großer Höhe manuell steuern. Da dies sehr selten vorkommt und nicht oft trainiert wird, fehlte es ihnen hierfür an Erfahrung (vgl. Abschnitt 1.1.2).

Des Weiteren weist Bainbridge (1983) darauf hin, dass ein Ausfall der Automation oft auch mit weiteren Anomalien und ungewöhnlichen Ereignissen und Situationen verbunden ist. So wurde der

Ausfall des Autopiloten bei dem beschriebenen Air-France-Unglück unter anderem auch durch einen Verlust reliabler Geschwindigkeitsinformation begleitet. In diesen Notfallsituationen sind besonders gute Kenntnisse und Fertigkeiten von Nöten, die ironischerweise durch die Automation verringert werden.

1.3 Lösungsansatz: adaptive Gestaltung der Mensch-Mensch-Maschine-Interaktion (MMI)

Den in den vorigen Abschnitten aufgeführten Problemen mit hochautomatisierten Systemen kann mit unterschiedlichen Maßnahmen begegnet werden. Neben Maßnahmen im Bereich Ausbildung und Training und im HMI (Human Machine Interface)-Design erwiesen sich Ansätze einer adaptiven Gestaltung der MMI (s. Abschnitt 2.1) als besonders vielversprechend. Das technische System reagiert dabei nicht starr wie bei konventionellen Automationskonzepten sondern kann sich flexibel an unterschiedliche Zustände oder Bedarfe des Interaktionspartners anpassen.

Adaptive Systeme sind darauf ausgelegt, den mentalen Zustand des Operateurs zu erfassen und Maßnahmen zu ergreifen, um kritischen Nutzerzuständen und damit einhergehenden Leistungseinbußen des Operateurs entgegenzuwirken. Wenn das System zum Beispiel erkennt, dass der Operateur aufgrund hoher Automation Schwierigkeiten hat, seine Aufmerksamkeit und sein Situationsbewusstsein aufrecht zu erhalten, könnte das technische System diesem Zustand entgegenwirken, indem es den Automationsgrad verringert (s. Abschnitt 2.1.2).

Ein Problem bei der Umsetzung adaptiver Technik besteht jedoch darin, dass in der realen Welt – wie der Unfallhergang von Flug AF447 deutlich macht – vielfältige Einflussfaktoren auf das Mensch-Maschine-System einwirken, die berücksichtigt werden sollten, damit das technische System effektiv unterstützen kann. In Laborumgebungen werden störende Einflussfaktoren jedoch in der Regel ausgeschaltet oder kontrolliert. Dies kann dazu führen, dass adaptive Systeme, die im Labor vielversprechende Ergebnisse erbringen, in der realen Welt nicht effektiv funktionieren, da sich Wechselwirkungen ergeben, die im Labor unzureichend berücksichtigt wurden. Möglicherweise konnten sich adaptive Systeme deshalb im operativen Kontext bislang noch nicht durchsetzen.

1.4 Eigenes Rahmenwerk adaptiver MMI

In dem Forschungsvorhaben, in das die vorliegende Dissertation eingebunden ist, wird daher eine ganzheitliche Betrachtung sowie Adressierung von Problemzuständen in adaptiver Mensch-Maschine-Interaktion angestrebt. Der ganzheitliche Ansatz zielt darauf ab, Faktoren, die auf Seiten des Operateurs, des technischen Systems und der Umwelt die Sicherheit und Effektivität des Mensch-Maschine-Systems beeinflussen und beeinträchtigen können, umfassend zu berücksichtigen. Problemzuständen soll dabei durch eine dynamische Auswahl und Anwendung von Unterstützungsstrategien situations- und bedarfsgerecht entgegengewirkt werden.

Vor diesem Hintergrund wurde im Rahmen des Forschungsvorhabens das in Abbildung 1 dargestellte Modell ganzheitlicher MMI konzipiert. Das Modell stellt in vereinfachter Form die grundsätzliche Funktionsweise des geplanten adaptiven technischen Systems sowie seine

Wechselwirkungen mit dem Operateur und der Umwelt dar. Damit bildet es den Ausgangspunkt für die Entwicklung einer multifaktoriellen Echtzeitdiagnose des Nutzerzustands im eigenen Promotionsvorhaben sowie für das zu entwickelnde dynamische Adaptierungsmanagement im Promotionsvorhaben von Sven Fuchs.

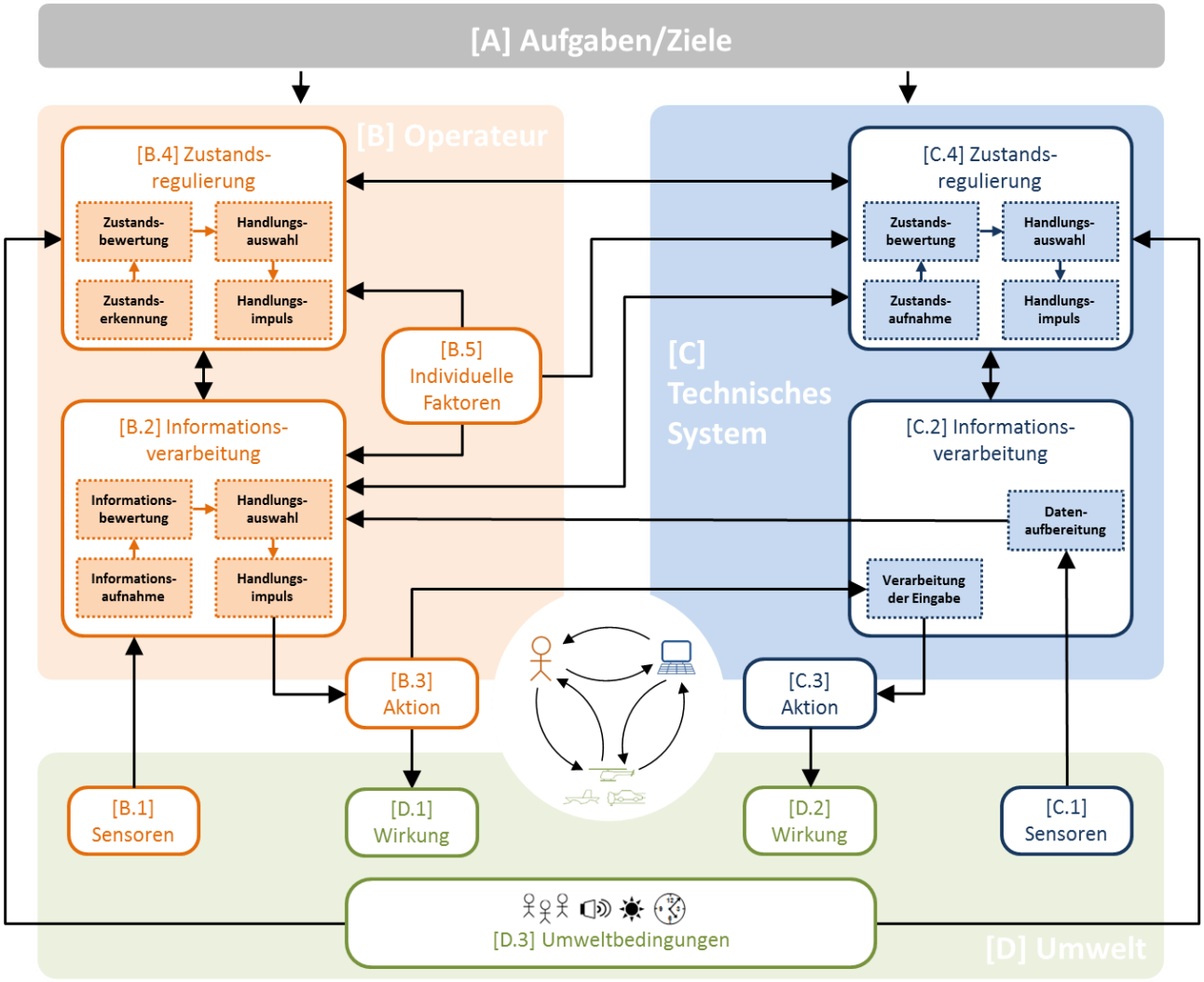


Abbildung 1. MMI-Modell mit aktiver Wechselwirkung zwischen Operateur und technischer Systemkomponente (Fuchs & Schwarz, 2014; Schwarz & Fuchs, 2014)

Das MMI-Modell in Abbildung 1 setzt sich zusammen aus den Aufgaben und Zielen des Mensch-Maschine-Systems (A) sowie dem Operateur (B), dem technischen System (C) und der Umwelt (D), die jeweils in Beziehung zueinander stehen (siehe Kreisdarstellung in der Mitte nach Flemisch et al., 2011). Das technische Subsystem verfügt über eine Komponente der Informations-verarbeitung (C.2), die auch in gegenwärtigen technischen Systemen ohne adaptive Systemfunktionalität enthalten ist. Diese Komponente hat die Aufgabe, über Sensoren (C.1) Daten aus der Umwelt zu erfassen und diese dem Operateur in aufbereiteter Form darzustellen (C.2). Zudem führt sie die vom Operateur über Bedieneingaben initiierten Aktionen aus (C.3). Zusätzlich verfügt das technische System noch über eine weitere Komponente, die im Modell als Zustandsregulierung (C.4) bezeichnet wird. Diese Komponente soll das technische System zu adaptivem Verhalten und einem aktiven Eingreifen in die Mensch-Maschine-Interaktion befähigen.

Wie in Abbildung 1 dargestellt ist, gliedert sich der Prozess der Zustandsregulierung des technischen Systems analog zu den Prozessen der Informationsverarbeitung (B.2) und Zustandsregulierung beim Menschen (B.4 – vgl. hierzu Abschnitt 3.4) in vier Stufen. Die ersten beiden Stufen „Zustandsaufnahme“ und „Zustandsbewertung“ beziehen sich auf die in der vorliegenden Dissertation erarbeitete Nutzerzustandsdiagnose, die den Namen *RASMUS* (Akronym für: Real time Assessment of Multidimensional User State) trägt (vgl. Schwarz & Fuchs, 2017). Die nachfolgenden Stufen „Handlungsauswahl“ und „Handlungsimpuls“ sind Teil des Adaptivitätsmanagements *ADAM* (Advanced Dynamic Adaptation Management; vgl. Fuchs & Schwarz, 2017). Eine in RASMUS vorgenommene ganzheitliche Erfassung und Bewertung von Umwelt- und Nutzerdaten dient dem Zweck zu bestimmen, (1) wann eine Adaptierung notwendig ist, und (2) welche möglichen Ursachen für Leistungsminderungen in Betracht kommen. ADAM nimmt auf dieser Grundlage die Auswahl, Konfiguration und Anwendung von Adaptierungsstrategien vor.

1.5 Zielsetzung des Promotionsvorhabens

Das Ziel der vorliegenden Arbeit bestand darin, ein Konzept für eine Nutzerzustandsdiagnose zu entwickeln, die im Sinne der im vorigen Abschnitt dargelegten ganzheitlichen Betrachtung eine umfassende Erfassung und Bewertung des mentalen Zustands eines Nutzers in adaptiver Mensch-Maschine-Interaktion ermöglicht. Die Diagnosekomponente RASMUS soll in der Lage sein, Problemzustände zu erkennen und Kontextinformationen bereit zu stellen, die das technische System für eine problemadäquate Auswahl und Anwendung von Unterstützungsstrategien benötigt. Um eine Übertragbarkeit auf unterschiedliche Anwendungsdomänen zu gewährleisten, soll das Diagnosekonzept generisch aufgebaut sein.

1.5.1 Anforderungen an das Diagnosekonzept

Der geplante Einsatz der Nutzerzustandsdiagnose im Bereich adaptiver Systemgestaltung ist insbesondere mit folgenden Anforderungen verbunden:

- Die Diagnose muss in (nahezu) Echtzeit erfolgen, damit das technische System frühzeitig auf Problemzustände reagieren kann.
- Der Nutzerzustand muss auf individueller Ebene ausgewertet werden (sogenannte „single trial analysis“), um auf Zustandsveränderungen eines einzelnen Operateurs reagieren zu können.

Der ganzheitliche Ansatz bringt zusätzliche Anforderungen an die Konzeption der Diagnose mit sich, durch die sich das Promotionsvorhaben von anderen Forschungsarbeiten zu diesem Thema unterscheidet. Die beiden Kernaspekte sind die multidimensionale Betrachtung und die multifaktorielle Bewertung des Nutzerzustands.

<u>Multidimensionale Betrachtung des Nutzerzustands</u>

Wie das Fallbeispiel in Abschnitt 1.1 illustriert, sind Leistungseinbußen in der realen Welt nicht zwingend das Resultat eines einzelnen Problemzustands, sondern entstehen möglicherweise erst

durch das Zusammenspiel verschiedener Einflussfaktoren und Problemzustände. Bei den Piloten von Flug AF447 waren dies:

- der emotionale Schock durch das Abschalten des Autopiloten,
- die Überforderung durch das manuelle Fliegen in Kombination mit der Fehlersuche,
- fehlerhaftes Situationsbewusstsein durch unzuverlässige Geschwindigkeitsangaben,
- sowie möglicherweise auch Aufmerksamkeitsprobleme durch das gleichzeitige Ertönen mehrerer Alarme.

In der vorliegenden Arbeit wird der Nutzerzustand daher *multidimensional* als Zusammenwirken von sechs mentalen Zuständen definiert (im Folgenden auch Dimensionen des Nutzerzustands genannt), die sich nachweislich positiv oder negativ auf die menschliche Leistungsfähigkeit auswirken können. Neben den im Fallbeispiel schon genannten Dimensionen (emotionaler Zustand, mentale Beanspruchung, Situationsbewusstsein, Aufmerksamkeit) zählen hierzu auch die Müdigkeit und die Motivation (siehe Kapitel 3 für nähere Ausführungen).

Multifaktorielle Erfassung und Bewertung des Nutzerzustands

Die Erfassung und Bewertung dieser Zustandsdimensionen soll *multifaktoriell* erfolgen. Dies bedeutet, dass in die Diagnose beispielsweise auch Umweltzustände, Aufgabenmerkmale und Ereignisse, die mit den betrachteten mentalen Zuständen in Zusammenhang stehen, einbezogen werden sollen. Ziel dieses Vorgehens ist es, das technische System dazu zu befähigen, nicht nur das Vorliegen kritischer Nutzerzustände zu erfassen, sondern auch mögliche Ursachen zu identifizieren, die im Rahmen der dynamischen Adaptierung adressiert werden können. Im Fall von Flug AF447 kann rückblickend zum Beispiel konstatiert werden, dass unter anderem das Abschalten des Autopiloten, unklare Rückmeldungen des technischen Systems und eine Vielzahl an Alarmen für kritische mentale Zustände der Piloten verantwortlich waren. Dabei handelt es sich um systemeigene Einflussfaktoren, die prinzipiell durch Adaptierung modifiziert werden könnten.

1.5.2 Forschungsfragen

Aus den Anforderungen ergeben sich einige Forschungsfragen, die bei der Konzeption einer multifaktoriellen Echtzeitdiagnose zu berücksichtigen sind. Tabelle 1 stellt dar, welche Forschungsfragen im Rahmen des Promotionsvorhabens untersucht wurden, und auf welche der in Abschnitt 1.5.1 aufgeführten Anforderungen diese Bezug nehmen.

Tabelle 1. Übersicht über die zu untersuchenden Forschungsfragen

Forschungsfrage	Bezug zu Anforderung
1. Welche Methoden eignen sich, um den Nutzerzustand multidimensional und in Echtzeit zu erfassen?	• Multidimensionale Betrachtung • Diagnose in Echtzeit
2. Welche Schwierigkeiten sind bei Anwendung der Methoden und Umsetzung einer Echtzeitdiagnose zu beachten?	• Diagnose in Echtzeit • Bewertung auf individueller Ebene
3. Wie wirken sich die verschiedenen Nutzerzustandsdimensionen auf die Leistung aus, und inwiefern beeinflussen sie sich gegenseitig?	• Multidimensionale Betrachtung
4. Welche weiteren Faktoren sind im Rahmen einer ganzheitlichen Betrachtung zu berücksichtigen?	• Multifaktorielle Bewertung
5. Wann weist ein Diagnoseergebnis auf eine „kritische“ Ausprägung des Nutzerzustands hin und signalisiert Unterstützungsbedarf?	• Diagnose in Echtzeit • Bewertung auf individueller Ebene

1.5.3 Validierung des Diagnosekonzepts

Des Weiteren war es das Ziel der Promotionsarbeit das generische Diagnosekonzept exemplarisch für einen Anwendungsfall aus der Domäne „Luftraumüberwachung“ umzusetzen und zu validieren. Hierbei ist zu beachten, dass die technische Umsetzung einer multifaktoriellen Echtzeitdiagnose für alle sechs betrachteten Zustandsdimensionen aufwändige Implementierungsarbeiten mit sich bringt, die nicht Bestandteil des Promotionsvorhabens sind. Im Sinne eines „Proof of concept“ sollte die Umsetzung daher zunächst auf drei Problemzustände beschränkt werden, die sich für den gewählten Anwendungsbereich als besonders relevant herausgestellt haben. Durch eine experimentelle Validierung der Diagnosefähigkeit sollte die Möglichkeit geschaffen werden, die Echtzeitdiagnose künftig als Grundlage für die Umsetzung von dynamischen Adaptierungsstrategien zu verwenden (vgl. Abschnitt 1.4).

1.6 Vorgehensweise und Gliederung der Dissertation

Die Vorgehensweise in der Dissertation ist in dem Arbeitsplan in Abbildung 2 veranschaulicht. Die Problemstellung und Zielsetzung bilden die Grundlage für die nachfolgenden Arbeitsschritte. Diese lassen sich thematisch in die drei Teilbereiche „theoretische Analyse“, „empirische Untersuchungen“ und „Entwicklung einer multifaktoriellen Echtzeitdiagnose“ unterteilen, welche in Abbildung 2 durch unterschiedliche Farben gekennzeichnet sind. Abschließend erfolgt eine allgemeine Bewertung der Forschungsarbeit. Während die theoretische Analyse die Grundlage für die anschließenden Arbeiten bildet, sind die empirischen Untersuchungen eng mit der Konzeption und Umsetzung der Echtzeitdiagnose verknüpft und dienen sowohl als Grundlage für die Entwicklung als auch zur Bewertung der multifaktoriellen Echtzeitdiagnose. Die Kapitel der Dissertation und die darin enthaltenen theoretischen, empirischen und konzeptionellen Arbeiten werden im Folgenden näher beschrieben.

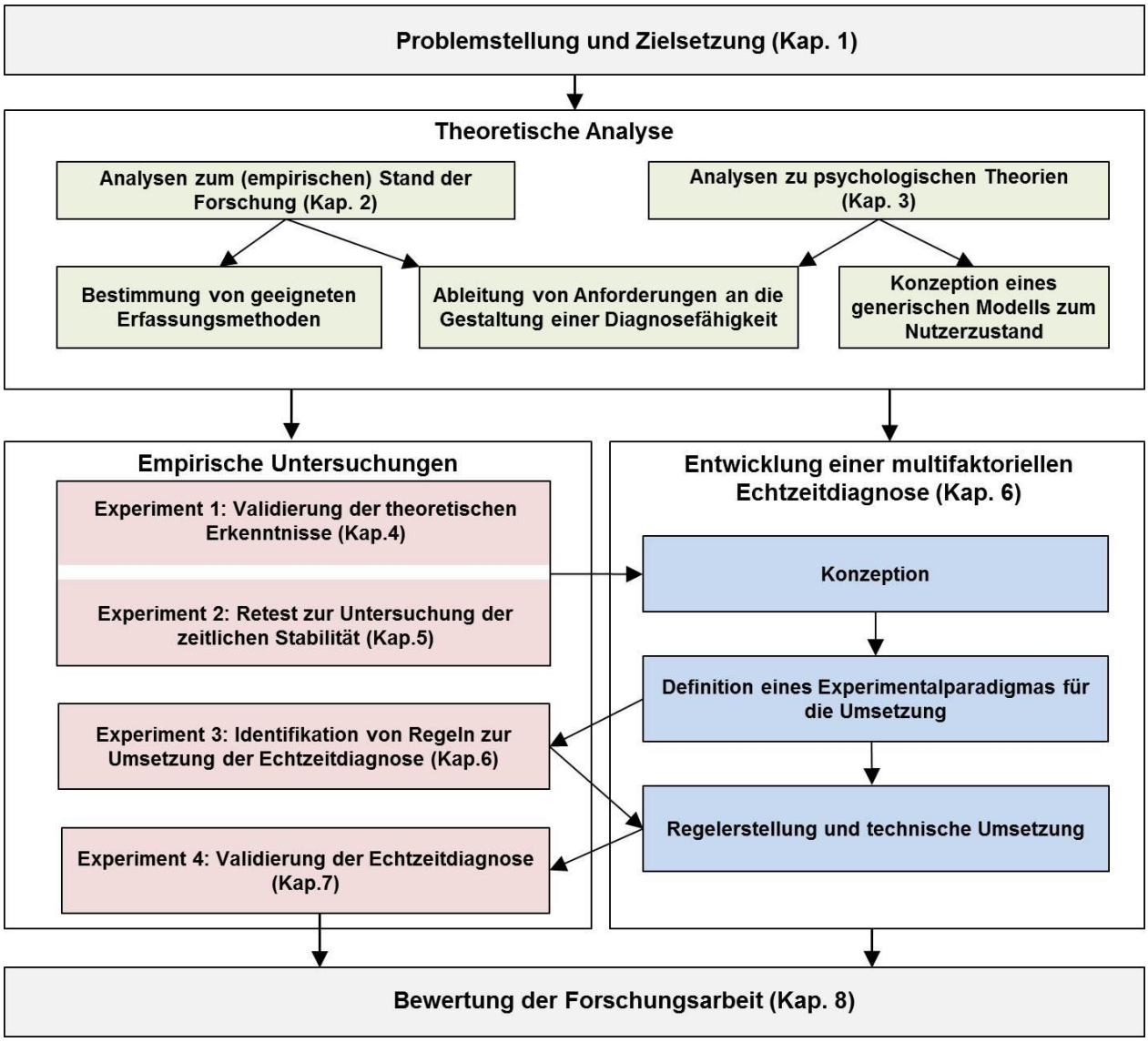


Abbildung 2. Arbeitsplan und Kapitelstruktur (Kapitelnummern in Klammern)

Auf Basis der Problemstellung und Zielsetzung (Kapitel 1) wurde zunächst eine Literaturanalyse durchgeführt, um Erkenntnisse zu den in Abschnitt 1.5.2 skizzierten Forschungsfragen zu gewinnen. Kapitel 2 befasst sich mit Analysen zum gegenwärtigen Stand der Forschung im Bereich adaptiver Systemgestaltung und Nutzerzustandserfassung. Ziel der Analysen war es, Methoden zu bestimmen, die geeignet sind, den Nutzerzustand multidimensional in einem adaptiven System zu erfassen und Aspekte zu identifizieren, die bei der Konzeption und Umsetzung einer Nutzerzustandsdiagnose in adaptiven Systemen zu berücksichtigen sind. Damit sollten insbesondere Erkenntnisse zu den Forschungsfragen 1 und 2 gewonnen werden (vgl. Tabelle 1). Kapitel 3 stellt die Erkenntnisse dar, die aus psychologischen Theorien und Modellen zu den betrachteten sechs Nutzerzustandsdimensionen abgeleitet wurden. Insbesondere sollten diese Analysen Erkenntnisse liefern, welche Wirkzusammenhänge zwischen den sechs Nutzerzustandsdimensionen und der Leistung bestehen, und welche Faktoren im Rahmen einer multifaktoriellen Bewertung des Nutzerzustands zu berücksichtigen sind. Diese Analysen adressieren somit die Forschungsfragen 3 und 4 in Tabelle 1. Die identifizierten Wirkzusammenhänge und Faktoren wurden anschließend in

ein generisches Modell zum Nutzerzustand überführt, das die theoretische Grundlage für die Konzeption einer multifaktoriellen Echtzeitdiagnose bildete.

Die Erkenntnisse aus den theoretischen Analysen wurden im nächsten Schritt empirisch überprüft. Das in Kapitel 4 beschriebene Experiment 1 zielte darauf ab, die Eignung verschiedener Diagnosemaße und Einflussfaktoren für eine multifaktorielle Nutzerzustandsbewertung zu untersuchen. Bei einem Retest (Experiment 2) wurde die zeitliche Stabilität dieser Befunde ein Jahr später geprüft (Kapitel 5). Da die Echtzeitdiagnose in der Lage sein muss, den Zustand eines einzelnen Nutzers zu erkennen, wurden diese Analysen nicht nur auf Gruppen- sondern auch auf Individualebene durchgeführt.

Kapitel 6 stellt sodann die Konzeption und Umsetzung einer multifaktoriellen Echtzeitdiagnose des Nutzerzustands vor, die auf den zuvor gewonnenen Erkenntnissen basiert. Hierbei ist anzumerken, dass für die Umsetzung der Echtzeitdiagnose eine andere Experimentalaufgabe verwendet wurde als bei den in den Kapiteln 4 und 5 beschriebenen Experimenten. Als Grundlage zur Bestimmung geeigneter Indikatoren und Regeln für kritische Indikatorausprägungen (vgl. Forschungsfrage 5 in Tabelle 1) wurde daher eine vorangegangene experimentelle Untersuchung (Experiment 3) herangezogen, die mit der gewählten Experimentalaufgabe durchgeführt worden war.

Kapitel 7 beschreibt anschließend das Validierungsexperiment (Experiment 4), in dem die Funktionalität und Genauigkeit der Echtzeitdiagnose in Hinblick auf drei ausgewählte Problemzustände experimentell untersucht wurde. Die Dissertation schließt mit einer Diskussion der gewonnenen Erkenntnisse und einem Ausblick auf potenzielle Folgeaktivitäten (Kapitel 8).

2 Stand der Forschung

Ansätze adaptiver Systemgestaltung werden schon seit einigen Jahrzehnten erforscht. In jüngerer Zeit konnte dieser Forschungsbereich von den technischen Fortschritten insbesondere im Bereich der Sensorik zur Nutzerzustandserfassung profitieren. So ergeben sich nun neue Möglichkeiten, mentale Zustände zu detektieren und durch adaptive Technik zu adressieren sowie diese Konzepte vom Labor in die praktische Anwendung zu übertragen. Bisherige Forschungsarbeiten geben dabei wertvolle Hinweise, welche Aspekte bei der Konzeption und Umsetzung adaptiver Systeme berücksichtigt werden sollten.

In diesem Kapitel wird zunächst ein kurzer Überblick über bisherige Ansätze einer adaptiven Gestaltung der Mensch-Maschine-Interaktion gegeben (Abschnitt 2.1). Die weiteren Abschnitte befassen sich mit dem gegenwärtigen Stand der Forschung zur Nutzerzustandserfassung im Kontext adaptiver Systemgestaltung. In Abschnitt 2.2 werden die Ergebnisse einer Literaturanalyse dargestellt. Sie geben Aufschluss darüber, inwiefern die sechs Dimensionen des Nutzerzustands, die im Fokus dieser Dissertation stehen (vgl. Abschnitt 1.5), in bisherigen Studien berücksichtigt wurden, und welche Verfahren und Ansätze für die Nutzerzustandsdiagnose in diesen Studien verwendet wurden. Anschließend erfolgt eine Bewertung der verschiedenen identifizierten Erfassungsmethoden in Hinblick auf ihre Eignung für eine multidimensionale Echtzeitdiagnose des Nutzerzustands (Abschnitt 2.3). Abschnitt 2.4 stellt abschließend dar, welche Erkenntnisse und Anforderungen für die Gestaltung einer Echtzeitdiagnose des Nutzerzustands aus diesen Analysen gewonnen werden konnten.

2.1 Bisherige Ansätze zur adaptiven Gestaltung der Mensch-Maschine-Interaktion

Die Forschung zu adaptiven Systemen begann bereits in den 70er Jahren (Rouse, 1976, 1977). Mittlerweile existiert ein breites Spektrum an Forschungsarbeiten zur adaptiven Systemgestaltung. Besonders einflussreich sind die Konzepte *Adaptive Aiding* (Rouse, 1988), *Adaptive Automation* (z.B. Parasuraman, Bahri, Deaton, Morrison, & Barnes, 1992; Scerbo, 1996) und *Augmented Cognition* (Schmorrow & Kruse, 2002). Diese werden in den nachfolgenden Abschnitten in kurzer Form vorgestellt.

2.1.1 Adaptive Aiding

Das Konzept Adaptive Aiding (vgl. Rouse, 1988) wurde zunächst für die Bereiche Flugzeugführung und Command & Control entwickelt und empirisch untersucht (z.B. Chu & Rouse, 1979; Morris & Rouse, 1986; Freedy, Madni, & Samet, 1985). Das Konzept sieht vor, im Regelfall dem Operateur die Ausführung der Aufgabe zu überlassen. So verbleibt dieser „im Loop“, also in den Regelkreis eingebunden. Der Operateur soll nur dann durch Automation unterstützt werden, wenn er diese Unterstützung benötigt, um die operationalen Anforderungen zu erfüllen. Dies beinhaltet, dass die Unterstützung zunehmen sollte, wenn die Aufgabenanforderungen so hoch werden, dass sich die Leistung des Operateurs ohne Unterstützung verschlechtern würde.

Rouse (1988) unterscheidet grundsätzlich zwischen drei Adaptierungsmethoden: Vereinfachung der Aufgabe („transformation“), Übernahme von Teilen der Aufgaben („partitioning“) und vollständige Übernahme einer Aufgabe („allocation“). Das Innovative zu der damaligen Zeit bestand darin, dass Initiierung und Auswahl der Unterstützung durch das technische System und nicht durch den Operateur erfolgen. Rouse (1981, 1988) argumentiert, dass hierdurch eine zusätzliche Belastung des Operateurs vermieden werden könne. Außerdem biete es die Möglichkeit, dass das technische System auch auf Problemsituationen, denen sich der Operateur nicht bewusst ist, reagieren kann.

Dieses Vorgehen setzt voraus, dass das technische System implizit die Leistung und den Unterstützungsbedarf des Operateurs bestimmen kann. Anfänglich wurden vorwiegend modellbasierte Vorhersagen verwendet, um die Leistung des Operateurs bei den gegebenen Aufgabenanforderungen zu bestimmen (Rouse, 1988). Es wurde zudem festgestellt, dass bestimmte Indikatoren, so genannte „leading indicators“, auf nahende Leistungsdefizite hinweisen (z.B. Zunahme der Reaktionszeit; Morris, Rouse, & Frey, 1985). Rouse (1981) schlägt außerdem die Erfassung und Analyse von Blickbewegungen des Operateurs vor, um Ablenkung oder eine zu starke Fokussierung der Aufmerksamkeit auf irrelevante Bereiche zu erkennen und diesen Problemzuständen entgegenzuwirken. Diese Methodik wurde in neueren Forschungsarbeiten umgesetzt (vgl. Abschnitt 2.2.5). Des Weiteren wurden verschiedene psychophysiologische Messmethoden z.B. von Wilson (2002) zur Bestimmung von Unterstützungsbedarf im Kontext von Adaptive Aiding herangezogen.

2.1.2 Adaptive Automation

Adaptive Automation zielt darauf ab, den in Abschnitt 1.2 genannten Automationsproblemen entgegenzuwirken, ohne auf die Vorteile von Automation verzichten zu müssen (Parasuraman et al., 1992; Scerbo, 1996). Gegenüber der konventionellen Automation, die statisch Aufgaben oder Aufgabenteile übernimmt, werden die Aufgaben bei Adaptiver Automation zwischen Operateur und Automation dynamisch umverteilt (vgl. *Dynamic Function Allocation*; Scerbo, 1996). Primäres Ziel adaptiver Automation ist die Regulierung der mentalen Beanspruchung: Bei hoher Beanspruchung soll das technische System den Operateur entlasten, indem es Aufgaben oder Aufgabenteile übernimmt. Bei geringer Beanspruchung soll es die Ausführung der Aufgabe(n) dem Operateur hingegen selbst überlassen, um Automationsprobleme (z.B. Unterforderung, Verlust des Situationsbewusstseins) zu vermeiden (vgl. Parasuraman et al., 1992). Um zu bestimmen, wann eine Umverteilung sinnvoll ist, werden unterschiedliche Methoden herangezogen. Viele Studien fokussieren sich auf (psycho-)physiologische Maße[1], die Aufschluss über den Grad der mentalen Beanspruchung geben können (z.B. Pope, Bogart, and Bartolome, 1995; Freeman, Milkulka, Prinzel, & Scerbo, 1999; Scerbo, Freeman, & Mikulka, 2000). Daneben wurden auch situative

[1] Einige Autoren verwenden den Begriff *psychophysiologische Maße*, um zu verdeutlichen, dass das physiologische Maß Aufschluss über den psychischen Zustand des Nutzers gibt. Zur einfacheren Handhabung wird in der vorliegenden Arbeit jedoch nur der Begriff der *physiologischen Maße* verwendet.

Anforderungen und Leistungsmaße herangezogen (z.B. Scallen & Hancock, 2001). Umfassender wird auf diese Arbeiten in Abschnitt 2.2 eingegangen.

Die Vorteile adaptiver Automation gegenüber konventioneller Automation konnten in verschiedenen Studien nachgewiesen werden (Kaber & Endsley, 2004; Hilburn, Byrne, & Parasuraman, 1997; Parasuraman, Mouloua, & Molloy 1996). Anwendungen in realen Systemen finden sich jedoch noch selten.

2.1.3 Augmented Cognition

Durch ein Forschungsprogramm der DARPA (Defense Advanced Research Projects Agency) hat sich in den USA Anfang dieses Jahrtausends das Forschungsfeld Augmented Cognition (kurz: AugCog) etabliert. Ziel der AugCog-Ansätze ist es, die menschliche Leistungsfähigkeit zu verbessern, indem mit physiologischen Methoden kognitive Problemzustände erfasst und diesen durch Adaptierung der Mensch-Maschine-Interaktion entgegengewirkt wird. Zumeist werden mehrere physiologische Maße für die Erfassung dieser Zustände herangezogen, z.B. Maße der Elektroenzephalografie (EEG), der Elektrokardiografie (EKG), der elektrodermalen Aktivität (EDA), des visuellen Systems (u.a. Pupillenweite, Fixationsdauer, Lidschlag), die durch Machine Learning-Verfahren (z.B. Artificial Neural Network; Bayes-Network) kombiniert ausgewertet werden (Stanney et al., 2009).

Im Unterschied zu den Konzepten der Adaptiven Automation beschränkt sich die AugCog-Forschung nicht auf die Erfassung und Adaptierung mentaler Beanspruchung. Vielmehr stehen „Flaschenhälse der menschlichen Informationsverarbeitung“ im Fokus, welche die Leistung des Menschen limitieren (Stanney et al., 2009). Diese beziehen sich insbesondere auf das sensorische System, das Arbeitsgedächtnis, die Ausführungsebene und die Aufmerksamkeit (Morrison, Kobus, & Brown, 2006). Validierungsexperimente im Rahmen des DARPA-Forschungsprogramms konnten zeigen, dass eine Detektion dieser Flaschenhälse und eine Adressierung durch Adaptierung möglich ist und mit Leistungsverbesserungen einhergeht (Tremoulet, Barton, & Craven, 2005; Morrison, Kobus, & Brown, 2006).

2.1.4 Resümee

Die bisherige Forschung zu adaptiven Systemen, die unter anderem im Rahmen von Adaptive Aiding, Adaptiver Automation und Augmented Cognition betrieben wurde, weist darauf hin, dass sich eine adaptive Gestaltung der Mensch-Maschine-Interaktion positiv auf die menschliche Leistungsfähigkeit auswirken und die Effektivität des gesamten Mensch-Maschine-Systems verbessern kann. Dennoch konnten sich adaptive Systeme im operativen Betrieb noch nicht durchsetzen. Dies kann vielfältige Gründe haben, wie mangelnde Akzeptanz der Operateure (Stuiver, Mulder, Brookhuis, de Waard, & Dijksterhuis, 2010; Menke, Best, Funke, & Strang, 2015) oder ungeklärte Fragen zur Verantwortlichkeit bei einem Versagen des Mensch-Maschine-Systems (Manzey, 2008). Für die Adaptivitätsforschung steht insbesondere die Frage im Vordergrund, wie die Genauigkeit und Robustheit adaptiver Systeme gesteigert werden kann. So können außerhalb von kontrollierten Laborumgebungen vielfältige Einflussfaktoren auf das Mensch-Maschine-System einwirken, die die Effizienz eines adaptiven Systems beeinträchtigen

können, wenn sie bei der Gestaltung der adaptiven Systemfunktionalität nicht berücksichtigt wurden. In den ersten AugCog-Studien zeigte sich zum Beispiel, dass es nicht ausreicht, Adaptierungen nur auf Basis der Detektion kognitiver Problemzustände vorzunehmen, sondern dass auch Kontextfaktoren berücksichtigt werden müssen (Stanney et al., 2009 mit Verweis auf eine Studie von Barker et al., 2004). Es setzt sich daher die Erkenntnis durch, dass zunächst solide Gestaltungsgrundlagen geschaffen werden müssen, bevor adaptive Systeme in Realumgebungen einsatzfähig sind (Steinhauser, Pavlas, & Hancock, 2009; Feigh, Dorneich, & Heyes, 2012).

2.2 Nutzerzustandserfassung in adaptiven Systemen

Aufgrund des generischen Ansatzes, der in dieser Dissertation verfolgt wird, erfolgten die Analysen zu bisherigen Studien der Nutzerzustandserfassung in adaptiven Systemen unabhängig von einer spezifischen Anwendungsdomäne. Tabelle 2 führt einige Forschungsarbeiten auf, in denen eine Nutzerzustandserfassung im Kontext adaptiver Systeme untersucht wurde. Sie stellt auch dar, welcher Anwendungsbereich und welche Dimension(en) des Nutzerzustands betrachtet wurden, sowie welche Methoden zur Diagnose des Nutzerzustands eingesetzt wurden. Die Anwendungsdomänen sind vielfältig: dazu zählen die Flugzeugführung, Flugsicherung, Kraftfahrzeugführung, militärische Anwendungen (z.B. im Bereich Command & Control - C2), die Prozessüberwachung, z.B. in Leitständen sowie Trainings- und Spieleanwendungen. Da viele dieser Arbeiten im Forschungsfeld der Adaptiven Automation durchgeführt wurden (vgl. Abschnitt 2.1.2), zeigt sich, dass sich die meisten Arbeiten auf die Erfassung der mentalen Beanspruchung beziehen. In den folgenden Abschnitten werden die eingesetzten Methoden und Ansätze für jeden Nutzerzustand in kurzer Form näher beschrieben.

Tabelle 2. Übersicht über betrachtete Nutzerzustände, Anwendungsbereiche und verwendete Methoden in Studien zu adaptiven Systemen

Nutzerzustand	Anwendungsbereich	Methode (Referenz)
Mentale Beanspruchung	Flugzeugführung	EEG, EOG (Belyavin, 2005)
		EDR freq., HR, HRV (Haarmann, Boucsein, & Schaefer, 2009)
		EEG, ECG (Dorneich et al., 2011)
		HRV und zerebralem Blutfluss (Parasuraman, 2003)
		HR, HRV, BP, BPV (Veltman & Jansen, 2003)
		EEG (Pope et al., 1995)
	Flugsicherung	Leistung in Zweitaufgabe (Kaber et al. 2002; Kaber & Wright, 2003)
		EEG, Herzrate, Lidschlag, Respiration (Wilson & Russell, 2003)
	Command & Control	HR, HRV, Respiration, Lidschlag (Veltman & Jansen, 2006)
		Pupillenweite, Fixationsdauer, Sakkadendistanz, Sakkadengeschwindigkeit (De Greef et al., 2009)
		objektorientiertes Aufgabenmodell und Leistung (Arciszewski, de Greef & van Delft 2009; DeGreef & Arciszewski 2009)
		Eyetracking-Maße (Van Orden, 2001)
		EEG, EKG, GSR (Tremoulet et al., 2005)
	Trackingaufgabe	EEG und ERP (Prinzel III, Pope, Freeman et al., 2001)
		ERP (Hadley et al., 1999; Scallen et al., 1995)

Nutzerzustand	Anwendungsbereich	Methode (Referenz)
	Operateure von UAVs	Kombination von EEG, ECG, EOG, EMG, Pupillengröße (Barker & Edwards, 2005)
	Kraftfahrzeugführung	Modellbasierte Bewertung (Hancock & Verwey, 1997)
	Prozessüberwachung	Kardiovaskuläre Maße, z.B. Blutdruck, Herzrate, HRV (Mulder et al., 2004)
Situations-bewusstsein	Command & Control	EEG, ERP (Berka et al., 2006; Fuchs et al. 2006)
Emotion	Flugzeugführung	HR + Gesichts-EMG für emotionalen Zustand; Aufgabenanalyse + Informationen zum persönlichen Hintergrund für Überzeugungen (Hudlicka & McNeese, 2002)
Aufmerksam-keit	kein spezifischer Anwendungsbereich	Ereignisse, Gesichtsausdruck, Theory of Mind (Bosse et al., 2008)
	Lernspiele	Gesichtsausdruck, Sitzposition, druckempfindliche Maus, Hautleitfähigkeit (Woolf et al., 2009)
	Command & Control	Eye-Tracking, Modellierung (Bosse, Lambalgen et al., 2009)
	Erkennen wichtiger Informationen in Texten und Bildern	Blickbewegung und EEG (Mathan et al., 2008; Behneman et al., 2009)
Müdigkeit	Kraftfahrzeugführung	EEG (Lin et al., 2006)
		Lidschlag, Blick- und Kopfbewegungen, Gesichtsausdruck (Ji, Zhu & Lan, 2004)
		Lidschlag und Sakkaden (Schleicher et al., 2008)
Motivation	Lernspiele	Interaktion mit dem System, Mausbewegungen (Ghergulescu & Muntean, 2010, 2011)
		Eye-Tracking (Blickabwendung) (D'Mello et al., 2012)
		EEG, GSR (Derbali & Frasson, 2010)
Mentale Beanspruchung, Situations-bewusstsein	Command & Control	Sitz- und Kopfposition (Balaban et al., 2005), Sitzposition (Frank, 2007)
Mentale Beanspruchung, Müdigkeit	Kraftfahrzeugführung	Fahrverhalten, Lidschlag, Kopfposition (Hancock & Verwey, 1997)
Operator Functional State	Kraftfahrzeugführung	43 physiologische Maße, u.a. EEG, Herzrate, Lidschlag (Wilson & Russell, 2003)
	Operateure von UAVs	EEG, EOG, ECG (Wilson & Russell, 2006)
	Command & Control	u.a. HRV, Atemfrequenz (Gagnon et al., 2014)
	Prozessüberwachung	Task Load Index bestehend aus EEG-basierten Maßen und HRV in Kombination mit Leistungsmaßen (Ting et al., 2008)

2.2.1 Erfassung mentaler Beanspruchung

Wie die Aufstellung in Tabelle 2 zeigt, wurden in den betrachteten Studien zum großen Teil physiologische Maße zur Erfassung der mentalen Beanspruchung eingesetzt. Dazu zählen kardiovaskuläre Maße, elektrodermale Maße, Maße der Hirnaktivität (insbesondere Elektro-enzephalographie – EEG) sowie okulomotorische Maße (vgl. Abschnitt 2.3 für eine Bewertung dieser Maße). Im Bereich der Kraftfahrzeugführung wurden auch verschiedene Leistungsmaße, die insbesondere das Lenkverhalten betreffen, herangezogen, um Rückschlüsse auf die Beanspruchung

zu ziehen (vgl. z.B. Son & Park, 2011; Hurwitz & Wheatley, 2002). Manche Studien legen auch eine modellbasierte Bewertung der Belastung zugrunde (Hancock & Verwey, 1997; Arciszewski, de Greef, & van Delft 2009).

Kombinierte Maße

Aus Tabelle 2 geht außerdem hervor, dass oftmals verschiedene physiologische Maße kombiniert wurden. Haarmann, Boucsein, & Schaefer (2009) kombinierten zum Beispiel die elektrodermale Reaktion (EDR) und die Herzrate (HR) bzw. Herzratenvariabilität (HRV) und stellten fest, dass die mentale Beanspruchung durch die Kombination genauer bestimmt werden kann als bei Verwendung der Einzelmaße. Weitere Studien berichten von einer Kombination der kardiovaskulären Maße HR und HRV (Mulder, Rusthoven, Kuperus, de Rivecourt, & de Waard 2007; de Rivecourt, Kuperus, Post, & Mulder, 2008), der okulomotorischen Maße Lidschlag-häufigkeit, Fixationsfrequenz und Pupillenweite (Van Orden, 2001) sowie einer Kombination von Lidschlag, Pupillenweite und Hauttemperatur im Gesicht (Wang, Duffy, & Du, 2007). Eine Kombination von Blickbewegungsmaßen mit EEG-basierten Maßen wird des Weiteren u.a. von Fuchs et al. (2006) vorgeschlagen, um diagnostizieren zu können, ob der Operateur kritische Ereignisse bewusst wahrgenommen hat (vgl. auch Abschnitt zur Erfassung und Bewertung des Situationsbewusstseins). Zudem gibt es Untersuchungen, in denen eine Vielzahl verschiedener physiologischer Parameter basierend auf EEG, EKG und okulomotorischen Daten mit Hilfe von Artificial Neural Networks bzw. Bayesian Networks zu Klassifikatoren zusammen geführt wurden. Beispiele sind die Untersuchungen von Barker & Edwards (2005), Wilson & Russell (2003, 2006), Van Orden (2001) und Lockheed Martin ATL (2005). Durch die Integration verschiedener Parameter durch Methoden der künstlichen Intelligenz konnten hohe Vorhersage-genauigkeiten von über 90 Prozent erzielt werden. Wilson & Russell (2003) berichten sogar von einer korrekten Unterscheidung zwischen einer Overload und Nicht-Overload-Bedingung von 97,5 Prozent.

Hybride Modelle

Neben der Kombination physiologischer Parameter wurden auch unterschiedliche Methoden-gruppen kombiniert (so genannte hybride Modelle). In der Literatur wird beispielsweise eine Kombination von Leistungsmaßen und aufgabenbezogenen Maßen (Scallen & Hancock, 2001, Hancock & Scallen, 1998, Parasuraman, Mouloua & Molloy, 1996) oder die Kombination kritischer Ereignisse mit Modellen zur Operateurleistung und dem aktuellen Zustand des Operateurs (Parasuraman et al., 1992) vorgeschlagen. Im Bereich Command & Control beschreiben de Greef & Arciszewski (2009) sowie Arciszewski, de Greef & van Delft (2009) einen Ansatz, bei dem ein objektorientiertes Aufgabenmodell in Verbindung mit einer Leistungsbewertung des Operateurs als Trigger für die Adaptierung zugrunde gelegt wird.

2.2.2 Erfassung des emotionalen Zustands

Der emotionale Zustand kann über subjektive, physiologische und verhaltensbasierte Maße erfasst werden (siehe auch Abschnitt 2.3). Bezogen auf das Verhalten äußern sich emotionale Zustände insbesondere in einem spezifischen Gesichtsausdruck. Zur Emotionserkennung werden daher häufig optische Analyseverfahren eingesetzt (z.B. Pantic & Rothkranz, 2003) oder Elektro-

myografie (EMG)-basierte Verfahren, welche die Muskelkontraktionen, die einen Gesichtsausdruck hervorrufen, analysieren (z.B. Mahlke & Minge, 2006). Eine weitere Möglichkeit besteht in der Analyse der Sprache, in der ebenfalls emotionale Zustände reflektiert sind (z.B. Lee & Narayanan, 2005; Nwe, Wie, & DeSilva, 2001). Busso et al. (2004) konnten empirisch zeigen, dass die Leistung und Robustheit eines Emotionserkennungssystems deutlich verbessert werden kann, wenn diese beiden Methoden kombiniert werden. Des Weiteren gibt es auch den Ansatz, emotionale Reaktionen über Eyetracking-Maße zu detektieren (de Lemos, Sadeghnia, Olafsdottir, & Jensen, 2008).

Der emotionale Zustand findet in der Forschung zu adaptiven Systemen in den letzten Jahren zunehmend an Beachtung. Beispielsweise berichten Hudlicka & McNeese (2002) von der Entwicklung des *ABAIS (Affect and Belief Adaptive Interface System)* für Air Force-Piloten. Das Interface passt sich dabei in Format und Inhalt an den affektiven Zustand des Nutzers, bestimmte Persönlichkeitseigenschaften und situationsspezifische Überzeugungen an, welche die Leistung beeinflussen können. Die Emotionserfassung erfolgt nach dem Modell von Russell (1980) über die Erfassung der beiden Komponenten Arousal und Valenz (vgl. Abschnitt 3.1.2). Für Arousal wurde die Herzrate verwendet, für die Valenz ein Gesichts-EMG. Das Konzept macht einen vielversprechenden Eindruck, scheint jedoch nicht weiterentwickelt worden zu sein.

In Bosse, Memon, & Treur (2008 und 2009) wird ein Modell beschrieben, mit dem die systemseitige Vorhersage emotionaler Zustände auf Basis der Erfassung und Bewertung von Ereignissen in der Umgebung und neuronalen Veränderungen, die eine Verhaltensreaktion auslösen (z.B. Gesichtsausdruck), möglich sein soll. Letzteres soll ermöglichen, auch dann einen Gefühlsausdruck zu detektieren, wenn dieser unterdrückt wird. Das Modell wurde bislang jedoch lediglich in Simulationsläufen validiert.

In einer anderen Studie im Bereich adaptiver Lernsysteme wurde der affektive Zustand (Zuversicht, Frustration, Langeweile, Aufregung, Interesse) durch die Kombination von vier verschiedenen Sensoren bewertet (Woolf et al., 2009). Verwendet wurde ein Gesichtserkennungssystem, die Analyse der Sitzposition über Drucksensoren, eine druckempfindliche Maus und ein Hautleitfähigkeitssensor. Die Autoren berichten, dass die Emotion *Frustration* mit einer Genauigkeit von 89 Prozent erfasst werden konnte, *Interesse* konnte die Maus als alleiniger Prädiktor mit einer Genauigkeit von 73 Prozent vorhersagen.

2.2.3 Erfassung der Motivation

Die Motivation wird als Trigger für eine Adaptierung im Bereich operationeller Systeme stark vernachlässigt. Berücksichtigung findet sie jedoch im Bereich adaptiver Lernsysteme. Eine Studie, die sich mit der Erfassung und Steigerung von Motivation bzw. Engagement bei Lernsystemen befasst, stammt von D'Mello, Olney, Williams, & Hays (2012). Hierbei wurde ein adaptives Lernsystem entwickelt, das Langeweile bzw. fehlendes Engagement der Schüler darüber detektiert, ob der Schüler auf den Bildschirm schaut. Wenn er dies nicht tat, wurde er vom System aufgefordert, wieder auf den Bildschirm zu sehen. Es konnten zwar Effekte auf die Lernleistung jedoch keine Verbesserung der Motivation festgestellt werden.

In anderen Studien wurde die Motivation bei Lernspielen über das Interaktionsverhalten mit dem System und den Fortschritt bei der Aufgabenbearbeitung erfasst (z.B. Ghergulescu & Muntean, 2010 und 2011). Zum Beispiel zeigte sich, dass bei hoher Motivation schneller von einer Aufgabe zur nächsten übergegangen wird (Touré-Tillery & Fishbach, 2014). In den meisten betrachteten Studien werden diese Maße im Rahmen eines modellbasierten Ansatzes ausgewertet und interpretiert (de Vicente & Pain, 2002; Ghergulescu & Muntean, 2000; Cetintas, Si, Xin, Hord, & Zhang, 2009).

Derbali & Frasson (2010) konnten außerdem zeigen, dass sich die Motivation auch über elektrophysiologische Maße erfassen lässt. Sie stellten fest, dass bei EEG-Messungen die Höhe der Theta-Welle im frontalen Bereich und der Beta-Welle im linken Zentrum positiv mit der Motivation korreliert. Zudem stellte sich in dieser Untersuchung die Hautleitfähigkeit als signifikanter Prädiktor der Motivation heraus.

2.2.4 Erfassung von Müdigkeit

Die Erfassung der Müdigkeit wurde in den letzten Jahren hauptsächlich im Automobilbereich erforscht (u.a. Schmidt, 2010; Sigari, 2009; May & Baldwin, 2009; Kecklund et al., 2007; Zhang & Zhang, 2006; Yuanyuan, 2006; Ji, Zhu, & Lan, 2004). Dabei wird zumeist das Ziel verfolgt, Konzepte zu entwickeln, wie müdigkeitsbedingten Fahrfehlern durch adaptive Fahrerassistenz-systeme entgegengewirkt werden kann. In einigen Studien wurde die Detektion von Müdigkeit auch im Bereich der Flugzeugführung untersucht (vgl. Review von Caldwell et al., 2009).

Die Erfassung der Müdigkeit kann ähnlich wie die mentale Beanspruchung über Leistungsmaße, subjektive Maße, physiologische und verhaltensbasierte Maße vorgenommen werden. Im Bereich der Kraftfahrzeugführung wurden zumeist okulomotorische Maße, insbesondere Maße des Lidschlagverhaltens, wie *PERCLOS (Percentage of eye closure*; Wierwille, Ellsworth, Wreggit, Fairbanks, & Kirn, 1994), Leistungsmaße oder eine Kombination dieser Methoden eingesetzt (vgl. z.B. Rauch, Kaussner, Krüger, Boverie, & Flemisch, 2009 sowie Reviews von Kecklund et al. 2007; Wright, Stone, Horberry, & Reed, 2007; Hartley, Horberry, Mabbott, & Krueger, 2000). Eine weitere Möglichkeit besteht darin, Müdigkeit kamerabasiert durch Analyse des Gesichtsausdrucks und der Kopfhaltung zu diagnostizieren (vgl. Ji, Zhu & Lan, 2004; Gu & Ji, 2004). Vergleichsweise selten werden EEG-basierte Maße im Bereich adaptiver Systemgestaltung eingesetzt (z.B. Lin et al., 2006), auch wenn sich diese als sensitiv und diagnostisch wertvolle Maße zur Erfassung der Müdigkeit erwiesen haben (vgl. Abschnitt 2.3.3).

2.2.5 Erfassung der Aufmerksamkeit und Vigilanz

Die Erfassung der visuellen Aufmerksamkeit erfolgt in der Regel über Fixationen und Blickbewegungen (z.B. Blair, Watson, Walshe, & Maj, 2009). Durch adaptive Systeme ist es möglich, die Aufmerksamkeit auf wichtige Stimuli zu lenken, die bislang nicht beachtet wurden. Bosse, van Lambalgen, van Maanen, & Treur (2009) setzten ein solches System für Überwachungs-aufgaben bei der Marine um. Anhand von Blickbewegungen ermittelt das System, worauf der Operateur seine Aufmerksamkeit richtet, vergleicht es mit einem präskriptiven Modell zur optimalen Aufmerksamkeitsverteilung und bietet bei Diskrepanzen Unterstützung an. Die Forscher

stellten in einer Validierungsstudie fest, dass eine Manipulation der Aufmerksamkeit durch Veränderung der Salienz von Objekten auf Basis der detektierten Aufmerksamkeit zu signifikant besseren Leistungen in einer Identifizierungsaufgabe führte.

Vigilanz wird auch als Daueraufmerksamkeit oder Wachsamkeit bezeichnet und wird speziell für Überwachungsaufgaben benötigt. Da sich Müdigkeit in der Regel in einer herabgesetzten Vigilanz äußert, sind die eingesetzten Erfassungsmethoden für Müdigkeit und Vigilanz meist dieselben. So lässt sich sowohl Müdigkeit als auch Vigilanz durch okulomotorische Maße, wie die Lidschlagfrequenz, die Lidschlagdauer und die Pupillenweite, erfassen (McIntire et al., 2011; McKinley, McIntire, Schmidt, Repperger, & Caldwell, 2011). Zudem wurde auch das PERCLOS-Maß erfolgreich eingesetzt, um die Aufmerksamkeit bei Vigilanzaufgaben zu bestimmen (Dinges & Grace, 1998).

2.2.6 Erfassung des Situationsbewusstseins

Die Methoden, mit denen das Situationsbewusstsein (engl. Situation Awareness – SA) erfasst werden kann, gliedern sich in drei Gruppen (vgl. Fracker, 1991; Uhlarik & Comerford, 2002):

- *explizite Verfahren*, die SA durch Abfrage von Informationen zur Situation erfassen, wie z.B. SAGAT (Endsley, 1988) und SPAM (Durso & Dattel, 2004),
- *implizite Verfahren*, in denen das Situationsbewusstsein indirekt über Leistungsparameter erfasst wird, wie z.B. das Erkennen kritischer Ereignisse (vgl. Fracker, 1991; Schwarz & Witt, 2011),
- *subjektive Verfahren*, die ein Selbst- oder Fremd-Rating beinhalten, z.B. SART (Taylor, 1990).

Es gibt jedoch wenige Studien, die sich damit beschäftigen, SA kontinuierlich mit Hilfe von physiologischen Sensoren zu erfassen. So stellt Wilson (2000) fest: *„Because essentially no research has been conducted to evaluate the utility of psychophysiological measures to determine an operator's level of SA, it is not clear what advantage these measures may have"*. Erst in den letzten Jahren wurden auch physiologische Methoden angewendet, um SA als Auslöser für adaptive Systeme zu verwenden. Studien, die einige Jahre nach der Feststellung von Wilson (2000) die Erfassung von SA mit physiologischen Methoden im Bereich Naval Command & Control untersuchen, werden in Fuchs et al. (2006) sowie Berka et al. (2006) berichtet. In diesen Studien wurde der Ansatz verfolgt, mit Hilfe von EEG- und ERP-Maßen vorherzusagen, ob der Operateur kritische Ereignisse erkennt oder nicht. Bereits Wilson (2000) wies darauf hin, dass das Auftreten von ERPs bei relevanten Stimuli dazu genutzt werden könnte, um zu bestimmen, ob ein Operateur den Stimulus sowie die Wichtigkeit dieses Stimulus erkannt hat. Berka et al. (2006) konnten zeigen, dass SA-relevante Ereignisse wie die korrekte / falsche Identifizierung von Tracks oder die Reaktionszeit auf sicherheitskritische Ereignisse durch ERPs bestimmt werden kann. Die Methode, Fixationen mit ERPs zu kombinieren, um festzustellen, ob eine bewusste oder unbewusste Wahrnehmung bestimmter Informationen oder Ereignisse stattfindet, wird auch als *EFRP (eye fixation related potentials*; Baccino et al., 2005) oder *FLERPs (fixation-locked ERP*; Hale et al., 2008) bezeichnet.

2.2.7 Resümee

In der Recherche zeigte sich, dass die meisten Arbeiten auf einen bestimmten Nutzerzustand, zumeist die mentale Beanspruchung, fokussiert sind. Deutlich seltener wurden andere Nutzerzustände, wie die Müdigkeit, Motivation, das Situationsbewusstsein, die Aufmerksamkeit oder emotionale Zustände für Adaptierungen zugrunde gelegt. Aus den Anwendungsbereichen geht hervor, dass Müdigkeit bislang im Bereich der Fahrzeugführung und Motivation ausschließlich im Bereich der Lernspiele und Tutoringsysteme betrachtet wurde. Dies legt nahe, dass domänenabhängig unterschiedliche Nutzerzustände als besonders relevant erachtet werden.

Insgesamt finden sich nur wenige Studien, die mehrere Nutzerzustände gleichzeitig berücksichtigen. Dem eigenen Ansatz einer multidimensionalen Betrachtung des Nutzerzustands am nächsten kommen Forschungsstudien, die sich auf das Konstrukt des *Operator Functional State (OFS)* beziehen (z.B. Hockey, 2003; Wilson et al., 2004). Hockey (2003) beschreibt OFS als ein *„multidimensional pattern of processes that mediate task performance under stress and high workload, in relation to task goals and their attendant physiological and psychological costs“* (Hockey, 2003 S.8). Das Konstrukt OFS beinhaltet somit ähnlich wie das Konzept des multidimensionalen Nutzerzustands eine mehrdimensionale Betrachtung von Prozessen und Faktoren, welche sich auf die Leistung bei der Aufgabenbearbeitung auswirken. Dabei werden, wie Hockey (2003) weiter ausführt, auch Umgebungsfaktoren, Eigenschaften sowie akute und chronische Zustände des Nutzers einbezogen, darunter Müdigkeit und Motivation. OFS fokussiert nach obiger Definition jedoch vordergründig auf die Zustände Stress und hohe Beanspruchung. In Forschungsstudien, die den OFS erfassen, wird dieses Konstrukt daher oft ähnlich operationalisiert wie die mentale Beanspruchung (vgl. z.B. Wilson & Russell, 2003).

In Hinblick auf die Erfassung der Nutzerzustände stellte sich in der Analyse heraus, dass die eingesetzten Methoden vielfältig sind. Besonders vielversprechend erscheinen Kombinationen verschiedener Maße und die Verwendung psychophysiologischer Maße. Im folgenden Abschnitt 2.3 werden die verschiedenen Methoden zur Erfassung der sechs Nutzerzustandsdimensionen näher analysiert und in Hinblick auf ihre Vor- und Nachteile für den Einsatz in der Echtzeitdiagnose RASMUS bewertet.

2.3 Bewertung der Methoden zur Nutzerzustandserfassung

Die in Abschnitt 2.3 aufgeführten Methoden zur Nutzerzustandserfassung und -bewertung lassen sich grundsätzlich den folgenden Methodengruppen zuordnen: Leistungsmaße, subjektive Verfahren, modellbasierte Bewertung sowie physiologische und verhaltensbasierte Maße. Die jeweiligen Vor- und Nachteile der Methoden sollen zunächst auf übergeordneter Ebene in Hinblick auf die genannten Methodengruppen analysiert werden (Abschnitt 2.3.2). Dabei werden die in Abschnitt 2.3.1 aufgeführten Bewertungskriterien zugrunde gelegt. Die Methodengruppe der physiologischen und verhaltensbasierten Maße, die auf Basis der Analysen am besten für eine Echtzeitdiagnose geeignet erscheint (vgl. Abschnitt 2.3.2), wird anschließend detaillierter beleuchtet (Abschnitt 2.3.3). Zum Ende dieses Abschnitts wird ein Resümee gezogen (Abschnitt 2.3.4).

2.3.1 Bewertungskriterien

Methoden zur Erfassung mentaler Zustände werden üblicherweise anhand von psychometrischen Gütekriterien, wie Validität, Reliabilität, Sensitivität und Diagnostizität bewertet. Im Kontext adaptiver Systemgestaltung sind zusätzlich auch anwendungsbezogene Aspekte wie die Praktikabilität und Interferenzfreiheit relevant (Kramer, 1991; Grandt, 2004). Bei der Praktikabilität wird bewertet, wie leicht die Methode angewendet werden kann und wie störanfällig die dabei eingesetzten Geräte sind. Zur Praktikabilität zählen auch Aspekte wie Kosten- und Trainingsaufwand. Interferenzfreiheit bezieht sich auf die Anforderung, dass die Erfassung den Nutzer bei der Aufgabenbearbeitung und in seinem Wohlbefinden nicht stören sollte. Als weitere Anforderungen für den Einsatz im Rahmen von RASMUS kommen hinzu, dass der Nutzerzustand kontinuierlich und in Echtzeit[2] erfasst und bewertet werden sollte, damit das technische System situations- und bedarfsgerechte Unterstützung anbieten kann (vgl. Abschnitt 1.5.1). Außerdem sollte die Methode (oder Methodenkombination) in der Lage sein, mehrere mentale Zustände simultan zu erfassen und zwischen diesen zu unterscheiden.

Für die Bewertung und Identifikation geeigneter Erfassungsmethoden sind somit folgende Bewertungskriterien relevant:

- Psychometrische Gütekriterien
- Praktikabilität
- Interferenzfreiheit
- Kontinuierliche Messung in Echtzeit
- Multidimensionale Erfassung

2.3.2 Bewertung der Methodengruppen

Subjektive Maße

Zu den subjektiven Maßen zählen Fragebögen, die den mentalen Zustand anhand einer Selbsteinschätzung des Nutzers (seltener auch über eine Fremdeinschätzung eines Beobachters) ein- oder mehrdimensional erfassen. Sie ermöglichen eine differenzierte Erfassung verschiedener Nutzerzustände. Zum Beispiel existieren Fragebögen zur Erfassung der mentalen Beanspruchung (z.B. *NASA-TLX*, Hart & Staveland, 1988; *RSME*, Zijstra, 1993), Müdigkeit (z.B. *Stanford Sleepiness Scale*, Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973; *Karolinska Sleepiness Scale*, Akerstedt, & Gillberg, 1990) oder emotionaler Zustände (z.B. *Self Assessment Manikin (SAM)*, Bradley & Lang, 1994). Subjektive Maße besitzen außerdem eine hohe Augenscheinvalidität (Cain, 2007). Ihre Eignung für adaptive Systeme ist jedoch dadurch eingeschränkt, dass der Nutzerzustand nur zu den jeweiligen Zeitunkten der Befragung erfasst wird, und die Befragung den Nutzer bei der Aufgabenbearbeitung zusätzlich beanspruchen, sowie stören oder ablenken

[2] Der Begriff „Echtzeit“ bezieht sich analog zu Hogervorst, Brouwer, & van Erp (2014) in der vorliegenden Arbeit auf eine Auswertung von Informationen innerhalb weniger Sekunden. Manche Autoren verwenden daher auch den Zusatz „nahezu“ Echtzeit.

kann. Des Weiteren sind Verzerrungen, z.B. aufgrund von sozialer Erwünschtheit und Antworttendenzen, möglich (Grandt, 2004, Dirican & Göktürk, 2011). Vermutlich wurden subjektive Maße deshalb nur selten für Adaptierungen herangezogen (vgl. Abschnitt 2.2). Da sie jedoch eine gezielte Erfassung unterschiedlicher Nutzerzustände ermöglichen und leicht anwendbar sind, werden sie in den Experimenten der vorliegenden Arbeit zur Validierung der verwendeten Diagnosemethoden eingesetzt.

Leistungsmaße

Durch Leistungsmaße wird die Leistung bei der Haupt- oder einer Nebenaufgabe bewertet, z.B. indem die Dauer bzw. die Reaktionszeit und die Fehlerrate erfasst werden. Leistungsmaße der Hauptaufgabe bieten den Vorteil, dass sie objektiv sind und nicht mit der Aufgabenbearbeitung interferieren. In einigen Anwendungsdomänen, wie der Fahrzeugführung, ist außerdem eine kontinuierliche Leistungserfassung, z.B. über das Lenkverhalten, möglich (vgl. z.B. Son & Park, 2011; Hurwitz & Wheatley, 2002). In anderen Anwendungsbereichen, insbesondere bei Überwachungstätigkeiten, wird eine kontinuierliche Erfassung jedoch dadurch erschwert, dass nur selten Aktionen des Operateurs erforderlich sind, anhand derer die Leistung bewertet werden kann. Das Zweitaufgabenparadigma sieht daher vor, die Leistung anhand einer Nebenaufgabe zu bewerten (vgl. Grandt, 2004). Dabei ist zu bedenken, dass Nebenaufgaben kognitive Ressourcen binden und mit der Bearbeitung der Hauptaufgabe interferieren können, insbesondere, wenn es sich um externe Nebenaufgaben handelt, die keinen Bezug zur Hauptaufgabe haben. In Hinblick auf die Nutzerzustandsbewertung in adaptiven Systemen besteht ein weiterer Nachteil von Leistungsmaßen in der mangelnden Diagnostizität. Das heißt, sie können keinen Aufschluss darüber geben, welche Nutzerzustände kritische Ausprägungen aufweisen.

Modellbasierte Bewertung

Neben diesen empirischen Erfassungsmethoden besteht auch die Möglichkeit durch eine modellbasierte Bewertung Prognosen über den kognitiven Zustand von Nutzern bei gegebenen Ausgangsbedingungen zu treffen. Hierfür stehen verschiedene Modellierungswerkzeuge zur Verfügung. Das Modellierungswerkzeug *IPME (Integrated Performance Modeling Environment)* der Firma *Micro Analysis and Design (MAAD)* bietet u.a. die Möglichkeit, die mentale Beanspruchung zu bewerten (Fowles-Winkler, 2003). Das Modellierungstool *FAST® (Fatigue Avoidance Scheduling Tool)* der Firma *Fatigue Science* ermöglicht eine Modellierung der Müdigkeit und der damit einhergehenden Leistungsfähigkeit auf Basis von Informationen zum Schlaf- / Wachrhythmus und der Schlafqualität. Die Modellierung ermöglicht es, Prognosen zur Leistungsfähigkeit und zu Veränderungen des Nutzerzustands zu treffen. Sie beruht jedoch auf Annahmen, die von den tatsächlichen Gegebenheiten im Einzelfall abweichen können (vgl. Parasuraman et al., 1992). In einer selbst durchgeführten Studie wurde die Leistungsfähigkeit von Operateuren in einem Schichtsystem durch FAST modelliert und mit der empirisch ermittelten Leistungsfähigkeit verglichen (Schwarz, Fuchs, Becker, & Kaster, 2013). Dabei kam es zu Abweichungen zwischen den realen Gegebenheiten und den Modellannahmen, was zu erheblichen Diskrepanzen in den Ergebnissen zwischen Empirie und Modell führte.

Physiologische und verhaltensbasierte Maße

Physiologische Maße beziehen sich auf Prozesse im menschlichen Körper, die durch Veränderungen des Nutzerzustands beeinflusst werden (z.B. Hirnaktivität, Herzrate, Pupillenweite, elektrodermale Aktivität) und kontinuierlich erfasst werden können. Verhaltensbasierte Maße (englisch: behavioral measures) werden unterschiedlich definiert. Wenn sie mit Leistungsmaßen, wie der Reaktionszeit, gleichgesetzt werden, wird argumentiert, dass sie im Gegensatz zu den physiologischen Maßen keine kontinuierliche Erfassung ermöglichen (z.B. Scerbo et al., 2001; Inagaki, 2003). In der vorliegenden Arbeit werden darunter jedoch, analog zu Elkin-Frankston, Bracken, Irvin, & Jenkins (2017), motorische Aktivitäten des Nutzers, wie Mimik, Gestik, Körperhaltung oder Bedieneingaben verstanden, die als Indikatoren für verschiedene Nutzerzustände herangezogen und zumeist kontinuierlich gemessen werden können. Nachteilig wirkt sich aus, dass die meisten physiologischen und verhaltensbasierten Maße auch durch andere Faktoren als den Nutzerzustand beeinflusst werden, so dass diese Faktoren entweder kontrolliert oder in die Diagnose mit einbezogen werden müssen (vgl. Abschnitt 2.4.1). Physiologische und verhaltensbasierte Maße werden üblicherweise über Sensoren erfasst (z.B. Eyetracker, Elektroden, Webcam). Dadurch interferiert die Messung nicht direkt mit der Bearbeitung der Hauptaufgabe. Das Fortschreiten der Technik ermöglicht es mittlerweile, dass der Nutzer durch die Sensoren kaum oder nur wenig beeinträchtigt wird.

Überblick über Vor- und Nachteile der Methodengruppen

Tabelle 3 fasst die wesentlichen Vor- und Nachteile der Methodengruppen für die Verwendung in RASMUS zusammen. Insgesamt kann konstatiert werden, dass subjektive Maße nicht zweckmäßig für eine Echtzeitdiagnose des Nutzerzustands sind, da eine regelmäßige Erfassung den Operateur bei der Aufgabenbearbeitung stören würde. Leistungsmaße sind insbesondere unter dem Aspekt der multidimensionalen Nutzerzustandserfassung ungünstig, da sie keinen Aufschluss darüber geben, welche Nutzerzustände bei Leistungsverschlechterungen beteiligt sind. Gegen die modellbasierte Bewertung ist einzuwenden, dass sie nur dann valide Aussagen zulässt, wenn die dem Modell zugrunde liegenden Annahmen mit der Realsituation übereinstimmen. In Anbetracht dessen, dass in der realen Welt vielfältige – auch unvorhersehbare – Einflussfaktoren wirken (vgl. Abschnitt 1.1), ist dies zu bezweifeln.

Physiologische und verhaltensbasierte Maße weisen das Problem auf, dass sie Störeinflüssen unterliegen und ihre Sensitivität, Reliabilität und zeitliche Stabilität daher situationsabhängig geprüft werden sollte. Im Vergleich zu den anderen Methoden überwiegen jedoch die Vorteile einer kontinuierlichen Messwerterfassung und Auswertung der Daten in Echtzeit, einer geringen Beeinträchtigung des Operateurs bei der Aufgabenbearbeitung und der Möglichkeit zur simultanen Erfassung verschiedener Nutzerzustandsdimensionen (Dirican & Göktürk, 2014; Grandt, 2004, Pfendler, Pitrella & Wiegand, 1995). Es verwundert somit nicht, dass sie besonders häufig zur Nutzerzustandserfassung in adaptiven Systemen herangezogen wurden (vgl. Abschnitt 2.2). Im folgenden Abschnitt wird daher die Eignung konkreter physiologischer und verhaltensbasierter Maße für die Nutzerzustandserfassung näher untersucht.

Tabelle 3. Übersicht über Vor- und Nachteile der Gruppen unterschiedlicher Erfassungsmethoden

Methode	Vorteile	Nachteile
Subjektive Maße	• Leicht anzuwenden • Sehr kostengünstig, da meist frei verfügbar • Validität und Reliabilität für viele Maße bestätigt	• Gefahr zusätzlicher Belastung und Ablenkung • Gefahr von Verzerrungen durch soziale Erwünschtheit • Keine kontinuierliche Erfassung ohne Beeinträchtigung des Operateurs
Leistungsmaße	• Leicht anzuwenden • Sehr kostengünstig • keine Beeinträchtigung des Operateurs • kontinuierliche Erfassung aufgabenabhängig möglich	• Keine Diagnostizität, daher keine multidimensionale Bewertung möglich • Bei Überwachungsaufgaben nicht kontinuierlich erfassbar
Modellbasierte Bewertung	• keine Beeinträchtigung des Operateurs • kontinuierliche Bewertung aufgaben- und modellabhängig möglich	• erfordert aufwändige Analysen • Validität fraglich (bietet nur Prognosen) • Kosten für Modellierungssoftware
Physiologische und verhaltensbasierte Maße	• keine oder geringe Beeinträchtigung des Operateurs • kontinuierliche Erfassung • teilweise hohe Diagnostizität – Diagnose verschiedener Nutzerzustände möglich	• Kosten für Sensoren • Störeinflüsse möglich, da diese Maße nicht nur durch mentale Prozesse beeinflusst werden • Sensitivität, Reliabilität und zeitliche Stabilität kann situationsabhängig variieren

2.3.3 Detailbewertung physiologischer und verhaltensbasierter Maße

Die Bewertung von physiologischen und verhaltensbasierten Maßen erfolgt gesondert für okulomotorische Maße bzw. Maße des visuellen Systems, die über einen Eyetracker erfasst werden können, Maße der Hirnaktivität (auch EEG-basierte Maße genannt), sowie weitere peripher-physiologische und verhaltensbasierte Maße, für deren Erfassung unterschiedliche Sensoren herangezogen werden.

Okulomotorische Maße

Mit Hilfe eines Eyetrackers können verschiedene okulomotorische Parameter, erfasst werden, wie Sakkaden, Fixationen, Lidschlüsse und die Pupillenweite. In einer vorangegangenen Analyse wurden verschiedene Messverfahren in Hinblick auf eine Erfassung der mentalen Beanspruchung und der Müdigkeit bei Radarbeobachtern untersucht (vgl. Schwarz, 2013; Schwarz et al., 2012). Die Verfahren wurden hinsichtlich ihrer psychometrischen Eigenschaften sowie in Bezug auf Anforderungen, die die Methode in Hinblick auf einen Einsatz im Realbetrieb erfüllen sollte (kontinuierliche Messwerterfassung, Bewertung des Nutzerzustandes in Echtzeit, Beeinträchtigungsfreiheit) bewertet. Die Analyse ergab, dass die Anforderungen am besten durch okulomotorische Maße, wie Fixationen, Lidschlag und Pupillenweite erfüllt werden. Der besondere Vorteil dieser Maße besteht darin, dass sie berührungslos und beeinträchtigungsfrei über so genannte Remote-Eyetracker erfasst werden können, die unterhalb des Monitors aufgestellt oder befestigt werden.

Tabelle 4 stellt wesentliche okulomotorische Parameter sowie Befunde zu ihrem Zusammenhang mit verschiedenen Dimensionen des Nutzerzustands dar. Es zeigt sich, dass viele okulomotorische Parameter als Indikatoren für mehrere Nutzerzustände herangezogen werden können. Während die

Maße bei Müdigkeit und Beanspruchung zumeist gegenläufig reagieren, weist z.B. die Pupillenweite gleich gerichtete Zusammenhänge mit der mentalen Beanspruchung, emotionalen Zuständen und Veränderungen der Aufmerksamkeit auf. Eine Differenzierung zwischen verschiedenen Nutzerzuständen auf Basis eines einzelnen Maßes erscheint somit nicht praktikabel.

Außerdem fällt auf, dass die Richtung des Zusammenhangs nicht nur zwischen den Maßen sondern auch innerhalb eines Maßes unterschiedlich ausfallen kann. So wurde festgestellt, dass Fixationen und Lidschlagparameter je nach Aufgabenart unterschiedlich reagieren. In Bezug auf den Lidschlag zeigte sich, dass dieser bei Wahrnehmungsaufgaben unterdrückt werden kann, um die Aufnahme visueller Informationen nicht zu behindern. Eine zunehmende Lidschlussfrequenz wurde hingegen bei kognitiven Aufgaben und sprachmotorischen Aktivitäten festgestellt (vgl. Review in Hargutt, 2003). Die Aufgabenart stellt somit einen weiteren zu berücksichtigenden Einflussfaktor dar.

Tabelle 4. Übersicht über okulomotorische Maße und ihren Zusammenhang mit verschiedenen Nutzerzuständen

Nutzerzustand	Indikator	Empirische Befunde
Aufmerksamkeit	Fixationen auf relevante Informationen	↑ - Blair et al., 2009; Bosse et al., 2009
	Lidschlussverhalten	↓ - Hargutt, 2003
	PERCLOS	↓ - Dinges & Grace, 1998
	Pupillenweite	↑ - wenn Fokus der Aufmerksamkeit wechselt; Laeng, Sirois, & Gredebäck, 2012
Mentale Beanspruchung	Fixationsdauer	↑ - bei kognitiver Beanspruchung; Pomplun, Reingold, & Shen, 2001; Meghanathan, Leeuwen, & Nikolaev, 2014 ↓ - bei visueller Beanspruchung; DeRivecourt et al., 2008; Duchowski, 2002
	Fixationsfrequenz	↑ - Van Orden et al., 2001; King, 2009
	Lidschlagfrequenz	↓ - bei visueller Beanspruchung; Veltman & Gaillard, 1996; Brookings, Wilson, & Swain, 1996 ↑ - bei kognitiver Beanspruchung; Wood & Hassett,1983; Tanaka & Yamaoka, 1993
	Nearest Neighbour Index (Verteilung der Fixationen)	↑ - Di Nocera et al., 2006
	Pupillenweite	↑ - Richer & Beatty, 1985; Van Orden et al., 2001; Iqbal et al., 2004; Wilson et al.,2004
	Index of Cognitive Activity (ICA)	↑ - Marshall, 2000; Schwalm, 2009
	Sakkadengeschwindigkeit	↑ - Di Stasi et al., 2010
	Sakkadenreichweite	↓- May et al., 1990
Emotionaler Zustand	Lidschlussfrequenz	↑ - bei Ärger / emotionaler Erregung; Ponder & Kennedy, 1927; Weiner & Concepcion, 1975
	Pupillenweite	↑ - bei emotionaler Erregung, Bradley et al., 2008
Motivation	Blickverhalten	Qu, Wang & Johnson, 2005
Müdigkeit	Fixationshäufigkeit	↓ - bzgl. Fixationen zwischen 150-900ms ↑ - bzgl. Fixationen <150 und >900ms; Wohleber et al., 2016
	Lidschlagamplitude	↓ - Caffier, Erdmann & Ullsperger, 2003
	Lidschlagdauer	↑ - Sirewaag & Stern, 2000
	Lidschlagfrequenz	↑ - Stern, Boyer & Schroeder, 1994; Stern et al., 1996
	Lidschlussdauer	↑ - Stern et al., 1996; Hargutt, 2003
	Lidschlussgeschwindigkeit	↓ - Ji et al., 2004
	PERCLOS	↑ - Barr, Popkin & Howarth, 2009
	Pupillenweite	↓ - Sirevaag & Stern, 2000
	Sakkadengeschwindigkeit	↓ - bei zunehmender Aufgabendauer; Sirevaag et al., 1999
	Sakkadenhäufigkeit	↓ - Stern et al., 1996

Anmerkung: ↑ Positiver Zusammenhang; ↓ Negativer Zusammenhang

Bei Fixationen zeigen sich ebenso unterschiedliche Ergebnisse, je nach dem, ob die Aufgabe vorwiegend visuelle oder kognitive Beanspruchung hervorruft. Außerdem scheint bei Blickbewegungen die räumliche Platzierung relevanter Informationen ein Einflussfaktor zu sein (Grandt, 2004). Bei der Interpretation dieser Maße besteht somit die Notwendigkeit aufgaben- und situationsspezifische Faktoren zu berücksichtigen.

Die Pupillenweite ist von der Beschaffenheit der Aufgabe weniger abhängig, reagiert jedoch unter anderem auf Lichtveränderungen (sog. Pupillenlichtreflex; Beatty & Lucero-Wagoner, 2000) und Koffeinkonsum (Stuiber, 2006). Insbesondere die Lichtverhältnisse können in vielen operativen Bedingungen z.B. der Fahrzeugführung nicht konstant gehalten werden, was die Anwendbarkeit der Pupillenweite einschränkt (Schwalm, 2009). Mit dem *Index of Cognitive Activity (ICA)* entwickelte Marshall (2000, 2002) ein Verfahren, das statt der absoluten Größenveränderung der Pupille, die durch Lichtverhältnisse beeinflusst wird, kurzfristige und schnelle Veränderungen der Pupillenweite heranzieht. Marshall stellte fest, dass diese hochfrequenten Komponenten der Pupillenweite auf Veränderungen der mentalen Beanspruchung sensitiv reagieren, aber von Lichteinflüssen unabhängig sind (Marshall, 2000, 2002). Das Verfahren setzt allerdings voraus, dass die Pupillenweite mit etwa 250 Hertz aufgezeichnet wird, was die Leistungsfähigkeit der meisten kommerziellen Eyetracker übersteigt.

Maße der Hirnaktivität

Elektroenzephalographie (EEG) bezeichnet die Messung elektrischer Hirnaktivität über die Kopfhaut. EEG-basierte Maße sind von hohem diagnostischem Wert, da sich zeigte, dass verschiedene kognitive Zustände mit spezifischen Aktivitätsmustern in den Frequenzbändern Alpha (8-13 Hz), Beta (>13 Hz), Theta (4-8 Hz) und Delta (< 4 Hz) verbunden sind. Darüber hinaus können *ereigniskorrelierte Potenziale (EKP bzw. engl.: event-related potentials, ERP),* insbesondere die P300-Komponente, Informationen über kognitive Reaktionen auf bestimmte Ereignisse liefern. In Bezug auf Aufmerksamkeit / Müdigkeit werden EEG-basierte Maße gelegentlich auch als „Goldstandard“ unter den physiologischen Maßen bezeichnet (Berka et al., 2005, Johnson et al., 2011). Neben EEG-basierten Maßen gibt es die Möglichkeit, kognitive Zustände insbesondere mentale Beanspruchung über die *funktionale Nah-Infrarot-Spektroskopie (fNIRS)* zu erfassen. Dabei handelt es sich um ein bildgebendes Verfahren, das mit der Hirnaktivität in Zusammenhang stehende Änderungen der optischen Eigenschaften des Hirngewebes erfasst (Böcker, 2014). Während EEG-basierte Verfahren eine hohe zeitliche, aber eine begrenzte räumliche Auflösung haben, weist fNIRS eine höhere räumliche aber geringere zeitliche Auflösung auf (Strait & Scheutz, 2014). Einige Forscher entwickelten daher kombinierte Anwendungen von EEG und fNIRS (z.B. Hirshfield et al., 2009; Nguyen, Ahn, Jang, Jun, & Kim, 2017), um die Vorteile beider Verfahren zu vereinen.

Tabelle 5 gibt einen Überblick über die zur Erfassung des Nutzerzustands eingesetzen Maße der Hirnaktivität und die empirischen Befunde. Es wird ersichtlich, dass Maße der Hirnaktivität Informationen über alle sechs betrachteten Zustandsdimensionen liefern können. Anwendung, Auswertung und Interpretation der Daten von EEG und fNIRS sind jedoch im Vergleich zu anderen physiologischen Maßen aufwändig und setzen spezielle Kenntnisse voraus. Unter anderem ist es notwendig, Artefakte herauszufiltern, die z.B. aufgrund von Augenbewegungen, Muskelaktivität

oder Bewegung auftreten können. Zur Artefaktkorrektur können verschiedene Signalverarbeitungstechniken eingesetzt werden (Wilson et al., 2004). Die Artefaktkorrektur und Auswertung wird bei fNIRS bislang vorwiegend offline, also nach der Datenaufzeichnung, vorgenommen (Strait & Scheutz, 2014), wobei in jüngerer Zeit auch Ansätze zur Echtzeiterfassung kognitiver Zustände mittels fNIRS entwickelt und untersucht wurden (Gateau, Durantin, Lancelot, Scannella, & Dehais, F., 2015; Hincks, Afergan, & Jacob, 2017).

In Hinblick auf EEG-basierte Verfahren sind mittlerweile sogenannte „low-cost EEG", verfügbar (z.B. Epoc EMOTIV, B-Alert X10, Quasar), die mit einer Analysesoftware ausgestattet sind, welche eine automatisierte Signalverarbeitung sowie Algorithmen zur Klassifizierung verschiedener kognitiver Zustände in Echtzeit anbieten (s. Tabelle 5). Vorteile bestehen insbesondere darin, dass die Headsets leicht anwendbar sind und für die Analyse und Interpretation der Klassifikatoren keine spezifischen Hintergrundkenntnisse notwendig sind. Für anwendungsnahe Forschungszwecke bieten EEG-Hersteller auch Headsets mit sogenannten trockenen Elektroden an, die statt eines leitenden Gels nur mit einer Kochsalzlösung befeuchtet werden und laut Herstellerangaben in wenigen Minuten angebracht und eingesetzt werden können (z.B. EMOTIV, 2013).

Tabelle 5. Übersicht über Maße der Hirnaktivität und ihren Zusammenhang mit verschiedenen Nutzerzuständen

Nutzerzustand	Indikator	Empirische Befunde
Aufmerksamkeit	Frequenz Alpha-Band	↓ - Makeig & Inlow, 1993
	Frequenz Theta-Band	↑ - Yamada, 1998 - reflektiert Interesse
	Klassifikator B-Alert Alertness	↑ - Berka et al., 2004
	Klassifikator Emotiv-Engagement	↑ - EMOTIV, 2013; Goldberg et al., 2011
	EKP (P300)	↑- Ververs Whitlow, Dorneich, & Mathan, 2005
Mentale Beanspruchung	Frequenz Alpha-Band	↓ - Barcelo, Gale & Hall, 1995; Brookings, Wilson & Swain, 1996; Berka et al., 2004
	Frequenz Beta-Band	↑ - Barcelo, Gale & Hall, 1995; Brookings, Wilson & Swain, 1996
	Frequenz Theta-Band	↑ - Brookings, Wilson, & Swain, 1996; Wilson, 2001
	Frequenzbänder kombiniert	↑ - Berka et al., 2007 – Mental Workload Index
	EKP (P300)	↓ - Kramer, 1991; Kok, 1997
	Klassifikator B-Alert	↑ - Berka et al. ,2004
	Klassifikator Emotiv Engagement	↑ - Wright, F. P., 2010
	fNIRS	- Gateau et al., 2015; Hincks, Afergan, & Jacob, 2016
Emotionaler Zustand	Aktivität in bestimmten Hirnegionen mit EEG bzw. fNIRS	Abhängig von der betrachteten Emotion - Mauss & Robinson, 2009; Tupak et al., 2014
	Klassifikator Emotiv Excitement	↑ - EMOTIV, 2013; Wright, 2010
	Klassifikator Emotiv Frustration	↑ - EMOTIV, 2013; Pröll, 2012; Goldberg et al., 2012
Motivation	Theta-Welle im frontalen Bereich	↑ - Derbali & Frasson, 2010
	Beta-Welle im linken Zentrum	↑ - Derbali & Frasson, 2010
Müdigkeit	Alpha-Aktivität	↑ - Kecklund et al., 2007
	Theta-Aktivität	↑ - Lal & Craig, 2001
	Klassifikator B-Alert Drowsiness	↑ - Johnson et al., 2011; Berka et al., 2005
	Kombination EEG/fNIRS	Nguyen et al., 2017
Situations-bewusstsein	EKP (P300)	- Wilson, 2000; Berka et al., 2006

Anmerkung: ↑ Positiver Zusammenhang; ↓ Negativer Zusammenhang

Peripherphysiologische und verhaltensbasierte Maße

Neben den okulomotorischen und EEG-basierten Maßen wurde in den in Abschnitt 2.2 betrachteten Forschungsstudien eine Vielzahl weiterer peripherphysiologischer und verhaltensbasierter Maße als Indikatoren für mentale Zustände im Kontext adaptiver Systemgestaltung eingesetzt. Als sensitiv für Veränderungen des Nutzerzustands stellten sich insbesondere Maße, die vom autonomen Nervensystem gesteuert werden, heraus (Grandt, 2004). Dazu zählen kardiovaskuläre Maße, wie die Herzrate, die Herzratenvariabilität (HRV) oder der Blutdruck, sowie Maße der Hautleitfähigkeit, der Körpertemperatur und der Atmung. Des Weiteren sind einige verhaltensbasierte Maße zu nennen, die unter anderem zur Erfassung des emotionalen Zustands, der Motivation und der Müdigkeit eingesetzt werden (z.B. Gesichtsausdruck, Sprache/Stimme, Kopfbewegungen). Eine Übersicht über diese Maße und ihren Zusammenhang mit dem Nutzerzustand ist in Tabelle 6 wiedergegeben.

Tabelle 6. Übersicht über peripherphysiologische und verhaltensbasierte Maße und ihren Zusammenhang mit verschiedenen Nutzerzuständen

Nutzerzustand	Indikator	Empirische Befunde
Mentale Beanspruchung	Atemfrequenz	↑ - Grassmann et al., 2016; Hogervorst et al., 2014; Veltman & Gaillard, 1998
	Blutdruck	↑ - Veltman & Gaillard, 1996
	Gesichtstemperatur	↓ - Or & Duffy, 2007; Veltman & Vos, 2005
	Hautleitfähigkeit	↑ - Reimer & Mehler, 2011; Roth, 1983
	Herzrate	↑ - Mulder et al., 2007; Roscoe, 1992; Boucsein & Backs, 2000
	HRV	↓ - Mulder et al., 2007; de Rivecourt et al., 2008
	Interbeatintervall	↓ - Veltman & Gaillard, 1996
Emotionaler Zustand	Atemfrequenz	↑ - bei negativen Emotionen – Boiten, Frijda, & Wientjes, 1994
	Blutdruck	↑ - bei negativen Emotionen höher als bei positiven Emotionen – Cacioppo et al., 2000
	Gesichtsausdruck	Indikator für Valenz; Russell, 1994; Pantic & Rothkranz, 2003
	Hautleitfähigkeit	↑ - bezogen auf das Arousal; Bradley & Lang, 2000
	Herzrate	↑ - bei negativen Emotionen höher als bei positiven Emotionen; Cacioppo et al., 2000
	HRV	↓ - bei Erschrecken; Ruiz-Padial, Sollers, Vila, & Thayer, 2003
	Muskelkontraktionen des Gesichts	Indikator für die Valenz - Mahlke & Minge, 2006
	Sprache (inhaltlich und akkustisch)	von der Emotion abhängig; Lee & Narayanan, 2005; Nwe, Wei & DeSilva, 2001
	Stimme z.B. Frequenz, Intensität,Geschwindigkeit	von der Emotion abhängig; Murray, & Arnott, 1993, Cowie et al., 2001; Johnstone & Scherer, 2000
	Verhalten: Mausdruck	Qi, Reynolds & Picard, 2001; Schaaff, Degen, Adler, & Adam (2012)
Motivation	Blickverhalten	Qu, Wang, & Johnson, 2005
	Hautleitfähigkeit	↓ - bezogen auf Selbstwirksamkeit; McQuiggan & Lester, 2006
	Herzrate	↓ - s.o.
	Geschwindigleit bei der Bearbeitung	↑- Touré-Tillery & Fishbach, 2014, de Vicente & Pain, 2002
	Mausbewegungen (zufällig/nicht zufällig)	de Vicente & Pain, 2002
Müdigkeit	Gesichtsausdruck	Ji, Zhu, & Lan, 2004; Gu & Ji, 2004
	Hautleitfähigkeit	↓ - Bundele & Banerjee, 2009; Yamamoto & Isshiki,1992
	Herzrate	↓ - Lal & Craig, 2000 ; Hankins & Wilson, 1998
	HRV	↑- Vicente et al., 2016; für Frequenzbereich 0,1 Hz - Egelund, 1982
	Kopfbewegungen	Mittal et al., 2016

Anmerkung: ↑ Positiver Zusammenhang; ↓ Negativer Zusammenhang

Peripherphysiologische und verhaltensbasierte Maße weisen gegenüber EEG-basierten Maßen den Vorteil auf, dass sie mit einem geringeren Mess- und Auswertungsaufwand verbunden sind. Allerdings spiegeln peripherphysiologische Maße den Erregungszustand (engl. Arousal) wider, der durch unterschiedliche Nutzerzustände beeinflusst werden kann (Hogervorst et al., 2014). Dementsprechend wurde in Hinblick auf die mentale Beanspruchung konstatiert, dass diese Maße nur über eine geringe Diagnostizität verfügen (Manzey, 1998).

Zu beachten ist auch die Beeinflussung der peripherphysiologischen Aktivität durch andere zustandsunabhängige Faktoren. Zum Beispiel reagiert die Herzrate auf körperliche Aktivität (Kay et al., 1975; Green et al., 1986), Koffeinkonsum (Barry et al., 2005) oder Änderungen der klimatischen Bedingungen (Grandt, 2004). Die HRV wird unter anderem durch die Körperposition, den circadianen Rhythmus, die Ernährung, das Alter und insbesondere auch durch die Atmung beeinflusst (Uhlig, 2018). Der Einfluss der Atmung auf die HRV wird auch als respiratorische Sinusarrhythmie (atemsynchrone Schwankung der Herzfrequenz) bezeichnet. Empfohlen wird daher, in experimentellen Untersuchungen zur HRV die Atemfrequenz und -tiefe mit zu erfassen, um Veränderungen der HRV korrekt zu interpretieren (Mulder, 1992). Da die Atmung durch Sprechen beeinflusst wird, wirkt sich dieses ebenfalls auf die HRV aus und sollte bei der Interpretation dieser Maße berücksichtigt werden (Veltman & Gaillard, 1998).

Zu den verhaltensbasierten Maßen zählen Bewertungen von Mimik und Gestik. Die Erfassung kann berührungslos über kamerabasierte Verfahren erfolgen. Eine andere Messmethodik stellt die Elektromyografie (EMG) des Gesichts dar, welche die Muskelkontraktionen, die einen Gesichtsausdruck hervorrufen, misst, und damit eine Bestimmung von kurzzeitigen und kaum wahrnehmbaren Veränderungen des Gesichtsausdrucks ermöglicht (Mahlke & Minge, 2006). Dies erfordert jedoch, ähnlich wie beim EEG, die Anbringung von Elektroden auf der Haut. Eine weitere Möglichkeit besteht in der Analyse der Sprache/Stimme, in der emotionale Zustände reflektiert sind (z.B. Lee & Narayanan, 2005; Nwe, Wei & DeSilva, 2001). Außerdem können auch Maße, die sich auf das Interaktionsverhalten mit dem technischen System beziehen, Aufschluss über den Nutzerzustand geben. Diese Maße haben den Vorteil, dass sie ohne die Erfordernis eines zusätzlichen Sensors über Logfiles des Systems ausgewertet werden können (Elkin-Frankston et al., 2017; Hershkovitz & Nachmias, 2011; Ben-Zadok et al., 2009).

2.3.4 Resümee

Wie die Literaturanalyse zeigt, existieren vielfältige Methoden zur Erfassung des Nutzerzustands insbesondere im Bereich physiologischer und verhaltensbasierter Messverfahren. Aufgrund der multidimensionalen und domänenunabhängigen Betrachtung konnten die Literaturbefunde für die einzelnen Zustandsdimensionen in diesem Abschnitt nur in knapper Form beschrieben und bewertet werden. Detailliertere Reviews der Methoden für spezifische Nutzerzustände finden sich für die *Müdigkeit* bei Barr, Popkin & Howarth (2009), Kecklund et al. (2007), Wright, Stone, Horberry, & Reed (2007) sowie Lal & Craig (2001), für die *mentale Beanspruchung* bei Kramer (1991), Miller (2001), Cain (2007) und Farmer & Brownson (2003), für *Aufmerksamkeit/Vigilanz* bei Oken, Salinsky, & Elsas (2006), für *emotionale Zustände* bei Mauss & Robinson (2009) und Calvo & D`Mello (2010) und für *Motivation* bei Touré-Tillery & Fishbach (2014).

Die Literaturanalyse konnte jedoch einige wesentliche Vor- und Nachteile der verschiedenen Methoden aufzeigen, auf deren Basis potenziell geeignete Erfassungsmethoden ausgewählt und näher untersucht werden sollen. Insgesamt weisen die Analyseergebnisse darauf hin, dass sich physiologische und verhaltensbasierte Maße besonders gut für eine Nutzerzustandserfassung in adaptiven Systemen eignen (vgl. Abschnitt 2.4.1).

In Bezug auf die Eignung konkreter physiologischer und verhaltensbasierter Maße ist festzustellen, dass okulomotorische Maße besonders positiv in Hinblick auf die Praktikabilität und Interferenzfreiheit bewertet werden können. Außerdem werden diese mit Fortschreiten der Technik immer kleiner und kostengünstiger. Ähnlich verhält es sich auch mit Sensoren zur Erfassung der Herzrate und Herzratenvariabilität, die in Fitnessarmbänder oder Smart Watches integriert sind und von vielen Menschen bereits im Alltag genutzt werden. Des Weiteren sind in dieser Hinsicht kamerabasierte Systeme vielversprechend, mit denen verschiedene verhaltensbasierte Parameter erfasst und softwaregestützt ausgewertet werden können.

EEG-basierte Maße weisen im Vergleich dazu eine höhere Intrusivität durch die Applikation von Elektroden auf der Kopfhaut auf. Außerdem ist der Kosten- und Zeitaufwand höher, auch wenn dieser durch die Einführung der Low-Cost-EEG bereits erheblich reduziert werden konnte. Ein bedeutender Vorteil ist jedoch die höhere Diagnostizität der EEG-basierten Maße, die eine exaktere Bestimmung spezifischer mentaler Zustände ermöglicht. Für den Anwendungsfall der multi-dimensionalen Nutzerzustandsdiagnose erscheinen insbesondere die systemeigenen Klassifikatoren vielversprechend, die von EEG-Herstellern zur Diagnose verschiedener mentaler Zustände angeboten werden. Sie sind über ein Software Development Kit (SDK) zugänglich und in nahezu Echtzeit auswertbar. Dabei ist allerdings zu bedenken, dass die zugrunde liegenden Auswertungsprozeduren in der Regel nicht bekannt gegeben werden, so dass die Funktionsweise der Klassifikation nicht nachvollzogen werden kann. Einige Forscher sprechen daher kritisch von so genannten „black box algorithms“ und raten von deren Anwendung ab (z.B. Fairclough, 2012).

Für das vorliegende Forschungsvorhaben wurde die Entscheidung getroffen, diese Klassifikatoren nicht von Vornherein für die Nutzerzustandserfassung auszuschließen, aber ihre Validität und zeitliche Stabilität für den konkreten Answendungsfall genau zu überprüfen. Auch für die übrigen Maße ist zu beachten, dass sich die Ergebnisse der betrachteten Forschungsstudien jeweils auf konkrete Anwendungsfälle und Ausprägungen eines Nutzerzustands beziehen. Die Analyse-ergebnisse werden daher für potenziell geeignete Maße noch einmal experimentell in Hinblick auf den eigenen Anwendungsfall überprüft. Die Untersuchungen von Maßen eines Low-cost-EEG, eines Eyetrackers und eines Multisensorgurtes werden in den Kapiteln 4 und 5 näher beschrieben und diskutiert.

In der Literaturanalyse wurde jedoch auch deutlich, dass die Verwendung von physiologischen und verhaltensbasierten Maßen zur Erfassung mentaler Zustände insbesondere im Realbetrieb mit verschiedenen Herausforderungen verbunden ist. Dazu zählt, dass sie Reaktionen des Menschen erfassen, die auch durch andere Faktoren als den mentalen Zustand beeinflusst werden. Beispiele hierfür wurden in den vorigen Abschnitten aufgeführt. In Laboruntersuchungen können diese Störeinflüsse konstant gehalten oder ausgeschaltet werden, was in realen Arbeitsumgebungen jedoch oft nicht möglich ist.

In diesem Zusammenhang ist auch zu beachten, dass viele dieser Maße sensitiv auf Veränderungen des Erregungszustands reagieren, der durch unterschiedliche Nutzerzustände beeinflusst werden kann. Beispielsweise können bestimmte emotionale Zustände, wie Angst oder Stress, die Atmungsfrequenz und den Blutdruck ebenso erhöhen wie hohe mentale Beanspruchung (vgl. Tabelle 6). Bei ausschließlicher Betrachtung dieser Maße kann somit nicht eindeutig zwischen diesen Zuständen differenziert werden. Ebenfalls bleiben die Ursachen für Veränderungen des Nutzerzustands unklar. In Abschnitt 2.4 wird näher darauf eingegangen, welche Implikationen dies für die Umsetzung der Echtzeitdiagnose in der vorliegenden Arbeit hat.

2.4 Erkenntnisse und Anforderungen für die Konzeption einer Echtzeitdiagnose

Die Forschungsstudien zu adaptiven Systemen und Methoden der Nutzerzustandserfassung wurden auch unter dem Aspekt analysiert, welche Herausforderungen bei der Umsetzung einer Echtzeitdiagnose bestehen, und welche Aspekte bei der Konzeption einer Diagnoseschnittstelle und der Anwendung von Erfassungsmethoden zu berücksichtigen sind. Einige Gestaltungsprobleme und -empfehlungen wurden bereits von anderen Autoren aufgezeigt und diskutiert (z.B. Steinhauser, Pavlas, & Hancock, 2009; Whitlow & Hayes, 2012; Feigh et al., 2012).

Tabelle 7 gibt eine Übersicht über die aus der Literaturanalyse extrahierten Erkenntnisse und stellt dar, welche Anforderungen sich daraus für die Konzeption der multifaktoriellen Echtzeitdiagnose ergeben. In den folgenden Abschnitten werden diese Aspekte näher erläutert.

Tabelle 7. Extrahierte Anforderungen an die Konzeption einer Echtzeitdiagnose

Nr.	Aspekt	Anforderung/Empfehlung
1	**Störeinflüsse**	• Eine Konfundierung der Diagnoseergebnisse durch Störeinflüsse sollte vermieden werden.
2	**Reliabilität und zeitliche Stabilität**	• Die zeitliche Stabilität von physiologischen Maßen sollte geprüft werden. • Verschiedene Erfassungsmethoden sollten kombiniert werden.
3	**Interindividuelle Unterschiede**	• Physiologische Maße sollten an einer Baseline relativiert werden. • Indikatoren sollten auf Individualebene in Hinblick auf ihre Eignung geprüft und nutzerspezifisch für die Diagnose ausgewählt werden.
4	**Menschliche Selbstregulierung**	• Zustände, in denen der Operateur sich nicht mehr selbst adaptieren kann, sollten diagnostiziert werden.
5	**Ursachenanalyse**	• Die Diagnose sollte bereits bei den Ursachen für kritische Nutzerzustände einsetzen.
6	**Kontextfaktoren**	• Aufgaben/Aufgabenstatus und Faktoren, die sich auf die Leistung des Operateurs auswirken, sollten identifiziert und erfasst werden.
7	**Oszillieren der Adaptierung**	• Es sollten Verfahren zur „Glättung“ der physiologischen Daten angewendet werden. • Aufgabenbeginn und -ende sollten identifiziert werden, um die Adaptierung darauf abstimmen zu können.
8	**Künstlichkeit von Laboraufgaben**	• Bei der experimentellen Untersuchung sollten realitätsnahe Aufgaben verwendet werden (z.B. Verwendung von Simulatoren).

2.4.1 Störeinflüsse

Bei der Analyse der Erfassungsmethoden zum Nutzerzustand zeigte sich, dass viele physiologische Maße durch verschiedene mentale Zustände beeinflusst werden. Dies ist bei der Auswahl der Erfassungsmethoden für eine multidimensionale Bewertung des Nutzerzustands zu beachten. Hinzu kommen weitere Faktoren, die nicht direkt mit dem Nutzerzustand in Zusammenhang stehen. Dazu zählen die Beschaffenheit der Aufgabe, Lichtverhältnisse, Koffeinkonsum, körperliche Aktivität oder Sprechen (vgl. Abschnitt 2.3.3). In Hinblick auf die Nutzerzustandsbewertung stellen diese Faktoren Störeinflüsse dar, da sie Veränderungen, die auf den Nutzerzustand zurückgehen, maskieren können (Kramer, 1991). In Laboruntersuchungen können diese Faktoren weitgehend ausgeschaltet oder konstant gehalten werden. Dies ist im operativen Betrieb jedoch nicht immer möglich. Es empfiehlt sich daher, bekannte Störeinflüsse mitzuerheben oder durch andere konzeptionelle Maßnahmen sicherzustellen, dass diese nicht zu fehlerhaften Diagnose-entscheidungen führen (Brouwer, Zander, van Erp, Korteling, & Bronkhorst, 2015).

2.4.2 Reliabilität und zeitliche Stabilität

Physiologische und verhaltensbasierte Maße können durch vielfältige bekannte und auch unbekannte Faktoren beeinflusst werden. Neben den im vorigen Punkt genannten Faktoren, die systematisch das Messergebnis beeinflussen, kann die Reliabilität und Stabilität auch durch Artefakte, variierende persönliche Stressfaktoren oder zeitliche Veränderungen eingeschränkt sein (Byrne & Parasuraman, 1996). Die Befunde von Faulstich et al. (1986) und Tomarken (1995) weisen darauf hin, dass die zeitliche Stabilität zwischen den Diagnosemaßen variiert (vgl. Abschnitt 5.1.1). Somit erscheint es ratsam, neben der Validität auch die zeitliche Stabilität physiologischer Maße für den geplanten Einsatzzweck zu überprüfen. Als Maßnahme, um eine robustere Diagnose zu erhalten, wurde, wie aus Abschnitt 2.1 hervorgeht, in den meisten Studien im Kontext adaptiver Systemgestaltung eine Kombination verschiedener Maße oder auch eine Kombination verschiedener Methoden im Rahmen von hybriden Modellen herangezogen. Kombinierte Maße bieten die Möglichkeit, dass Störeinflüsse, die bei einzelnen Maßen auftreten, durch andere Maße kompensiert werden können, so dass die Robustheit und Reliabilität der Diagnose deutlich gesteigert werden kann (Stanney et al., 2009). Dies konnte auch experimentell nachgewiesen werden (z.B. Wilson & Russell, 2003; Haarmann et al., 2009).

2.4.3 Interindividuelle Unterschiede

Bei Verwendung von physiologischen und verhaltensbasierten Maßen muss berücksichtigt werden, dass sich Personen auch unabhängig vom mentalen Zustand in ihren physiologischen und verhaltensbasierten Reaktionen unterscheiden können. Beispiele sind anatomisch bedingte Unterschiede in der Pupillengröße, Unterschiede im (Ruhe-)Puls oder der Lidschlagfrequenz. Physiologische Messwerte werden daher üblicherweise an einer individuellen Baseline (Messwerte im Ruhezustand) relativiert (Scerbo et al., 2001).

Des Weiteren ist es notwendig, dass das technische System Unterschiede im mentalen Zustand eines einzelnen Operateurs während der Laufzeit diagnostizieren (so genannte „Single-Trial“-Diagnose, vgl. Anforderung in Abschnitt 1.5.1) und auf diese reagieren kann (vgl. Mulder et al.

2008). Auswertungen auf Gruppenebene, wie sie üblicherweise in experimentellen Untersuchungen vorgenommen werden, sind hierbei nicht aussagekräftig. Zum einen kann die Sensitivität physiologischer Maße für Veränderungen des Nutzerzustands interindividuell unterschiedlich ausfallen (Veltman & Jansen, 2003). Zum anderen gehen Personen auch unterschiedlich mit bestimmten Anforderungen um, was interindividuelle Unterschiede in der Beanspruchung und Leistung zur Folge haben kann (vgl. Belyavin 2005, Parasuraman et al., 1992; Veltman & Jansen, 2003). Es ist somit möglich, dass ein Parameter zwar auf Gruppenebene statistisch signifikant Veränderungen des Nutzerzustands anzeigt, auf Individualebene aber nicht sensitiv reagiert (vgl. Hogervorst, Brouwer, & van Erp, 2014). Die Eignung der Diagnosemaße muss daher auf Individualebene geprüft werden. Veltman & Jansen (2003) schlagen dabei vor, für jede Person zu ermitteln, welche Maße am aussagekräftigsten sind, sowie welche Werte für diese eine kritische Ausprägung des Nutzerzustands (hohe und niedrige Beanspruchung) repräsentieren.

2.4.4 Menschliche Selbstregulierung

Bei der Gestaltung adaptiver Systeme ist zu berücksichtigen, dass es sich bei dem menschlichen Organismus selbst um ein adaptives System handelt (vgl. Hancock & Chignell, 1987; Veltman & Jansen, 2003). Dies bedeutet, dass sich der Mensch kontinuierlich an sich verändernde Arbeitsanforderungen anpasst. Physiologische Reaktionen, die z.B. auf eine erhöhte Anstrengung hinweisen, sind ein Zeichen des adaptiven Verhaltens Die Tatsache, dass sich der Operateur anstrengt, bedeutet jedoch nicht zwangsläufig, dass er auch Unterstützung benötigt. Nach Veltman & Jansen (2003) spiegelt hohe mentale Anstrengung auch eine hohe Involviertheit des Operateurs in die Aufgabe wider. In diesem Fall würde sich die Leistung verschlechtern, wenn aufgrund der gestiegenen Anstrengung des Operateurs ein höherer Automationsgrad zugewiesen wird. Veltman & Jansen (2006) empfehlen daher, dass ein adaptives System in normalen Situationen, in denen der Operateur durch zusätzliche Anstrengung die Aufgaben bewältigen kann, keine Aufgaben abnehmen/verändern sollte. Physiologische Maße könnten allerdings nützlich sein, um anzuzeigen, wenn der Operateur so stark überlastet ist, dass er sich nicht mehr adäquat adaptiert. Eine andere Strategie des Operateurs mit nicht erfüllbaren Aufgabenanforderungen umzugehen, kann neben der Aufbringung von Anstrengung auch darin bestehen, die Ziele der Aufgabenbearbeitung anzupassen (s. Abschnitte 3.1.3 und 3.4). Dies könnte sich beispielsweise darin äußern, dass bestimmte Aufgaben oder Aufgabenteile ignoriert werden. In diesem Fall wäre die Anwendung einer Adaptierungsstrategie notwendig, die den Operateur zu einer angemessenen Aufgabenbewältigung befähigt.

2.4.5 Ursachenanalyse

Sciarini & Nicholson (2009) weisen darauf hin, dass Stress und Müdigkeit nur Reaktionen und keine Indikatoren der kognitiven Leistungsfähigkeit sind. Wichtig sei es, die Mechanismen zu verstehen, die diese Reaktionen erklären. Gleichermaßen stellen auch Steinhauser et al. (2009) heraus, dass eine Adaptierung auf Basis von Symptomen nicht ausreichend sei, um Situationen, in denen der Operateur nicht mehr leistungsfähig ist, frühzeitig entgegenwirken zu können. Die Diagnose des Nutzerzustands sollte somit nicht nur erfassen, wenn der Mensch nicht mehr leistungsfähig ist, sondern auch mögliche Ursachen für kritische Zustände analysieren. Wie sich

zeigte, können physiologische und verhaltensbasierte Maße zwar (mit unterschiedlicher Spezifität) Aufschluss über Veränderungen mentaler Zustände geben (vgl. Abschnitt 2.3.3). Die Ursachen für Veränderungen des Nutzerzustands bleiben dabei jedoch unklar. Somit ist eine zusätzliche Erfassung von Faktoren, die kritischen Nutzerzuständen zugrunde liegen, erforderlich.

2.4.6 Kontextfaktoren

Verschiedene Forscher weisen darauf hin, dass neben dem Nutzerzustand auch der Kontext und die Umweltbedingungen sowie die individuellen Eigenschaften des Operateurs betrachtet werden müssen (vgl. beispielsweise Whitlow & Hayes, 2012; Steinhauser et al., 2009; Fuchs et al., 2006). Nach Fuchs et al. (2006) werden Kontextinformationen und Aufgabenstatus insbesondere dazu benötigt, um nicht nur entscheiden zu können, wann adaptiert werden sollte, sondern auch was und wie zu adaptieren ist. Zu berücksichtigen ist auch, dass eine Nichtbeachtung des Aufgabenstatus dazu führen kann, dass eine Adaptierung zu ungünstigen Zeiten an- und abgestellt wird (vgl. Forderung bzgl. Oszillation der Adaptierung).

Feigh et al. (2012) führen eine Klassifizierung verschiedener Faktoren auf, die neben dem Nutzerzustand als Auslöser für die Adaptierung eingesetzt werden können. Unterschieden wird dabei zwischen systembasierten, umweltbasierten, aufgaben- und missionsbasierten sowie räumlich-zeitlichen Triggern. Eine eigene Analyse von relevanten Kontextfaktoren wird in Kapitel 3 dargestellt.

2.4.7 Oszillieren der Adaptierung

Bei Verwendung physiologischer Sensoren wurde festgestellt, dass diese oft eine hohe Änderungssensitivität aufweisen, was dazu führt, dass die aufgezeichneten Werte oszillieren und der Grenzwert, ab dem die Adaptierung einsetzt, mehrmals in kurzer Zeit über- und unterschritten wird (Barker, Edwards, O'Neill, & Tollar, 2004; Inagaki, 2003). Es zeigte sich, dass dieses Oszillieren – auch „Yo-yoing" genannt (Stanney et al., 2009; Diethe, 2005) – unerwünschte Effekte auf die Leistung des Operateurs hat, da es ihn verwirren und seine mentale Beanspruchung erhöhen kann (Stanney et al., 2009). Um dem Oszillieren entgegenzuwirken, wird vorgeschlagen, die physiologischen Messungen zu glätten, z.B. durch die Anwendung von Filtern (Diethe et al., 2004) oder die Daten über Zeitintervalle zu mitteln (Roscoe, 1993; Haarmann et al, 2009). Eine weitere Möglichkeit besteht darin, ein Zeitfenster zu definieren, in dem die Adaptierung nach der Auslösung bestehen bleiben muss, bevor diese ausgeschaltet werden kann. Barker et al. (2004) verwendeten hierfür ein zehnsekündiges Zeitfenster (sog. „Deadband"). Diethe et al. (2004) führten eine so genannte „refractory period" ein, die ähnlich wie das Deadband bei Barker et al. (2004) ein Zeitintervall definiert, in dem Zustandsveränderungen bestehen bleiben müssen.

Zu berücksichtigen ist auch, dass ein Oszillieren der Adaptierung in einem Regelkreissystem auch durch die Adaptierung selbst zustande kommen kann: Wenn die mentale Beanspruchung hoch ist, setzt die Adaptierung ein; durch die Adaptierung verringert sich die Beanspruchung, bis diese den kritischen Grenzwert unterschreitet und die Adaptierung aussetzt. Dies kann dazu führen, dass die Beanspruchung wieder steigt und die Adaptierung wieder neu einsetzen muss (vgl. Stanney et al., 2009). Um diesem Problem zu entgehen, schlagen Barker et al. (2004) vor, die Adaptierung erst

auszusetzen, wenn eine Aufgabe, die zu hoher Beanspruchung geführt hat, abgeschlossen ist. Eine solche Strategie wird auch in Dorneich et al. (2004) angewendet, setzt aber voraus, dass das System erkennt, welche Aufgabe der Operateur gerade bearbeitet und wann diese beendet ist.

2.4.8 Künstlichkeit von Laboraufgaben

In experimentellen Untersuchungen werden oft künstliche Aufgaben verwendet, um gezielt bestimmte mentale Zustände hervorzurufen und untersuchen zu können (z.B. Oddball-Aufgabe, n-back Test). Eine Übertragbarkeit der Erkenntnisse auf operative Bedingungen ist jedoch eingeschränkt, da sie die Dynamik und das Aufgabenspektrum in der Praxis nicht wiedergeben und oft von vergleichsweise kurzer Dauer sind. Mulder et al. (2008) weisen darauf hin, dass für die Untersuchung von Zustandsänderungen, die erst nach längerer Zeit auftreten, wie dies z.B. bei Müdigkeit der Fall ist, (semi-)realistische Arbeitsbedingungen hergestellt werden sollten. Weitgehend realistische Arbeitsbedingungen können durch Simulatoren hergestellt werden. Beispielsweise werden im Automobilbereich Fahrsimulatoren und im Bereich der Flugzeugführung Cockpit-Simulatoren eingesetzt, in denen die üblichen operativen Aufgaben realitätsnah und ohne Risiko für die Umwelt ausgeführt werden können.

3 Theoretische Grundlagen für eine multifaktorielle Nutzerzustandsdiagnose

Wie das zu Beginn beschriebene Unglück von Flug AF447 verdeutlichte, können verschiedene mentale Problemzustände zu Leistungseinbußen und Fehlverhalten führen (vgl. Abschnitt 1.3). Der in der Dissertation verfolgte Ansatz einer multifaktoriellen Diagnose des Nutzerzustands sieht daher vor, den Nutzerzustand – im Unterschied zu der oft eindimensionalen Betrachtung in anderen Forschungsstudien (vgl. Abschnitt 2.2) – als ein multidimensionales Konstrukt zu verstehen. Anhand einer Literaturanalyse wurden sechs mentale Zustände identifiziert, die sich nachweislich positiv oder negativ auf die menschliche Leistungsfähigkeit auswirken können (vgl. Tabelle 8). Der Nutzerzustand wird in der vorliegenden Arbeit somit multidimensional als ein Zusammenwirken dieser sechs Zustandsdimensionen definiert.

Tabelle 8. Übersicht über die sechs in der Dissertation betrachteten Nutzerzustände und ihre Auswirkungen auf die Leistung

Nutzerzustand	Auswirkungen auf die Leistung	Referenz
Mentale Beanspruchung	Leistung kann sich verschlechtern, wenn die Beanspruchung zu hoch oder zu gering ist.	Hancock & Chignell, 1987,1988
Emotionaler Zustand	Bestimmte emotionale Zustände (z.B. Angst, Stress) können die Aufmerksamkeit einengen und die Informationsverarbeitung beeinträchtigen.	Staal, 2004
Motivation	Motivation kann die Leistungsfähigkeit verbessern und eine Zeit lang auch bei Schlafentzug aufrechthalten.	Wilkinson et al.,1966
Müdigkeit	Müdigkeit trägt zu 20-30% der Unfälle im Transportbereich bei.	Akerstedt et al., 2003
Aufmerksamkeit	Aufmerksamkeitsdefizite (z.B. attentional tunneling) und Vigilanzverlust bei dauerhaften Überwachungsaufgaben beeinträchtigen die Leistung.	Wickens, 2005 Mackworth, 1948
Situations-bewusstsein	Ein Großteil menschlicher Fehler (bei Flugzeugunfällen 88%) ist auf Probleme mit dem Situationsbewusstsein zurückzuführen.	Endsley, 1999

Auf die in Tabelle 8 aufgeführten Auswirkungen der mentalen Zustände auf die Leistung wird in der folgenden Analyse von psychologischen Theorien und Modellen zu den einzelnen Nutzerzustandsdimensionen näher eingegangen (Abschnitt 3.1). Ziel dieser Analyse ist es, Erkenntnisse zu relevanten Einflussfaktoren, Unterscheidungen und Wirkzusammenhängen zu gewinnen, die in der Nutzerzustandsdiagnose berücksichtigt werden sollten (vgl. Anforderungen 5 und 6 in Tabelle 7). Im Anschluss werden identifizierte Wechselwirkungen zwischen den sechs Zustandsdimensionen (Abschnitt 3.2) sowie Einflussfaktoren (Abschnitt 3.3) detaillierter beleuchtet. Die Erkenntnisse werden daraufhin in ein allgemeines Modell zum Nutzerzustand integriert (vgl. Abschnitt 3.4). Das Modell stellt die wesentlichen Wirkfaktoren im Zusammenhang mit den sechs Zustandsdimensionen dar und bildet die Grundlage für die Konzeption der multifaktoriellen Echtzeitdiagnose.

3.1 Theorien und Modelle zu den sechs Dimensionen des Nutzerzustands

Im Folgenden werden einige ausgewählte Theorien und Modelle zu den sechs Dimensionen des Nutzerzustands näher ausgeführt, die Aufschluss über wichtige Unterscheidungen, Einflussfaktoren und Wirkzusammenhänge geben. Die wesentlichen Erkenntnisse, die daraus für die Gestaltung adaptiver Systeme folgen, werden am Ende eines jeden Abschnitts zusammengefasst.

3.1.1 Mentale Beanspruchung

Zur mentalen Beanspruchung existiert eine Vielzahl an Theorien und Modellen. Der Fokus in den folgenden Ausführungen liegt daher auf den theoretischen Erkenntnissen, die im Kontext adaptiver Systemgestaltung von besonderer praktischer Relevanz sind.

Klassifikation der Beanspruchung

Die Beanspruchung ist ein multidimensionales Konstrukt, das sich in verschiedene Arten oder Ausprägungsformen untergliedern lässt. Zunächst kann zwischen *psychischer und physischer Beanspruchung* unterschieden werden, wobei sich psychische Beanspruchung auf kognitive Vorgänge bezieht und physische Beanspruchung aus körperlich anstrengenden Belastungen resultiert.

Die psychische Beanspruchung lässt sich weiter unterteilen in *mentale und emotionale Beanspruchung*. Während mentale Beanspruchung durch *aufgabenspezifische* Belastungsfaktoren, wie Schwierigkeit und Komplexität, bestimmt wird, bezieht sich emotionale Beanspruchung auf Stresszustände, die durch *ausführungsspezifische* Belastungsfaktoren (z.B. Zeitdruck, Lärm, Hitze, Gefahren, soziale Konflikte) ausgelöst werden und mit aversiven Gefühlen wie Angst und Hilflosigkeit verbunden sind (Ribback, 2003). Die mentale Beanspruchung kann sich auf unterschiedliche mentale Vorgänge beziehen und lässt sich somit weiter in *perzeptive, kognitive und psychomotorische Beanspruchungsformen* unterteilen (vgl. Ausführungen zur Multiple Ressource Theory weiter unten).

Im Folgenden liegt der Fokus auf der psychischen Beanspruchung und hierbei insbesondere auf der mentalen Beanspruchung, da Tätigkeiten in den für die adaptive Systemgestaltung relevanten Domänen (vgl. Abschnitt 2.2) vorwiegend kognitive Anforderungen beinhalten. Der Aspekt der emotionalen Beanspruchung in Form von Stress wird in den Analysen zum emotionalen Zustand betrachtet (vgl. Abschnitt 3.1.2).

Definition psychischer Beanspruchung

Eine Definition der psychischen Beanspruchung findet sich in der ISO-Norm DIN EN 10 075-1 (2000). Die Definition basiert auf dem Belastungs-Beanspruchungskonzept von Rohmert (1984), das zwischen psychischer *Belastung* und psychischer *Beanspruchung* unterscheidet. Unter psychischer Belastung wird die „Gesamtheit aller erfassbaren Einflüsse, die von außen auf den Menschen zukommen und psychisch auf ihn einwirken“ (DIN EN 10 075-1, 2000) verstanden. Im Arbeitskontext resultieren diese Einflüsse aus den Arbeitsbedingungen. Hierzu zählen Art und Umfang der Arbeitsaufgabe, die Arbeitsmittel (alle technischen Geräte am Arbeitsplatz), die

Arbeitsumgebung (z.B. Beleuchtung, Schall, Klima), die Arbeitsorganisation (z.B. Arbeitszeiten, Arbeitsabläufe) und der Arbeitsplatz (direkte Arbeitsumgebung).

Psychische Beanspruchung ist ein Resultat der psychischen Belastung und wird definiert als „die unmittelbare (nicht langfristige) Auswirkung der psychischen Belastung im Individuum in Abhängigkeit von seinen jeweiligen überdauernden und augenblicklichen Voraussetzungen, einschließlich der individuellen Bewältigungsstrategien“ (DIN EN 10 075-1, 2000). Psychische Beanspruchung hängt somit einerseits von den äußeren Belastungsfaktoren ab, wird aber auch durch überdauernde und augenblickliche Merkmale, Eigenschaften und Verhaltensweisen des Menschen beeinflusst. Die jeweiligen Einflussfaktoren sind in Anlehnung an eine Darstellung von Hilburn & Jorna (2001) in Abbildung 3 veranschaulicht. Da die ursprüngliche Darstellung nur überdauernde Faktoren aufführt, wurde die Darstellung entsprechend der Definition nach DIN EN 10 075-1 (2000) um die augenblicklichen Voraussetzungen (variable Faktoren) ergänzt.

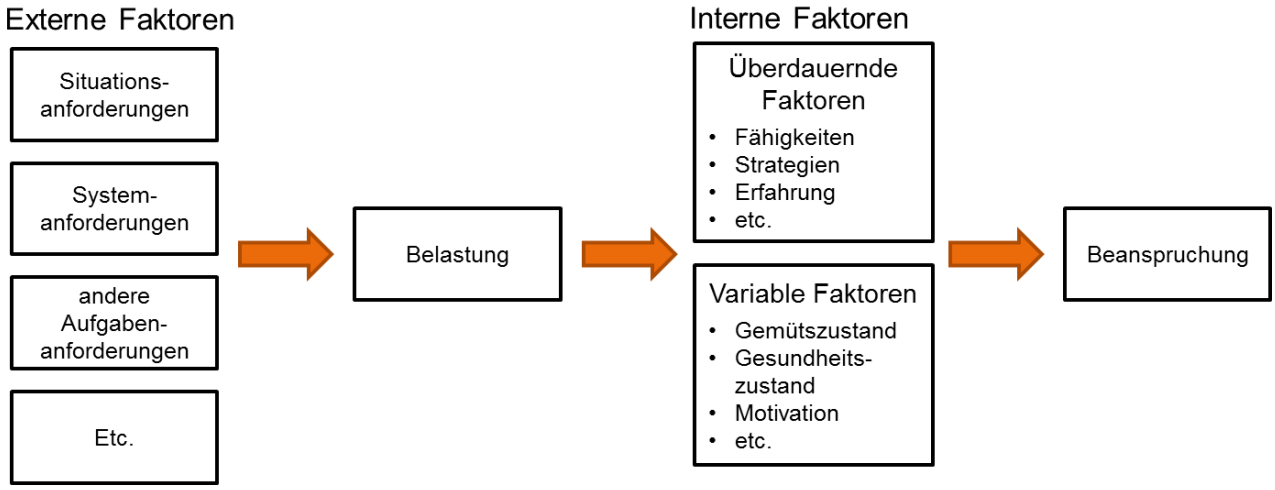


Abbildung 3. Einflussfaktoren auf die psychische Beanspruchung nach DIN EN 10 075-1 (2000)

Zu den relativ überdauernden Merkmalen zählen unter anderem Persönlichkeitseigenschaften, Fähigkeiten, Fertigkeiten, Kenntnisse und Erfahrungen. Zu den augenblicklichen Merkmalen lassen sich beispielsweise die Müdigkeit, die Gemütslage und die Motivation zählen. Die psychische Beanspruchung kann somit bei gleichbleibenden Belastungsfaktoren sowohl zwischen verschiedenen Personen als auch bei derselben Person zu verschiedenen Zeitpunkten variieren. Zum Beispiel führt eine gewöhnliche Autofahrt bei einem erfahrenen Autofahrer in der Regel zu einer geringen Beanspruchung, während ein Führerscheinneuling sehr stark beansprucht sein kann. Ebenso kann die gleiche Autofahrt auch bei einem geübten Autofahrer zu einer hohen Beanspruchung führen, wenn dieser abgelenkt ist, müde ist oder sich schlecht fühlt. Die Darstellung macht deutlich, dass eine Vielzahl unterschiedlicher Faktoren zusammenwirken, die bei einer Person zu einem bestimmten Grad psychischer Beanspruchung führen.

Multiple Resource Theory

Wickens (1984) unterscheidet in seiner *Multiple Resource Theory* unterschiedliche Ressourcen-dimensionen, auf die sich mentale Beanspruchung beziehen kann und stellt diese in einem Würfel-modell dar (vgl. Abbildung 4). Neben den Dimensionen Verarbeitungsstufe, Wahrnehmungs-

modalität, Kodierung /Ausführungsart, nimmt Wickens (2002) später noch die Dimension focal vs. ambient hinzu, die sich auf die visuelle Informationsaufnahme bezieht.

Die Informationsverarbeitung gliedert sich in dem Würfelmodell in die drei Stufen: *Perzeption*, *Kognition* und *Ausführung*. Wickens (2002) stellte durch Untersuchungen mit Zweifachaufgaben fest, dass perzeptive und kognitive Aufgaben jeweils die gleiche Ressource beanspruchen, während die Handlungsausführung auf eine andere Ressource zurückgreift. Dies führt dazu, dass die Ausführung einer perzeptiven oder kognitiven Aufgabe nicht mit der parallelen Ausführung einer motorischen Aufgabe interferiert, während die parallele Ausführung einer sensorischen und einer kognitiven Aufgabe zu Leistungseinbußen bei einer Aufgabe führt.

Neben den Stufen der Informationsverarbeitung differenziert Wickens (1984, 2002) zwischen den unterschiedlichen Modalitäten *visuell* und *auditiv* sowie zwischen der *räumlichen* und *verbalen* Kodierung einer Information. Er stellte fest, dass Informationen, die, auch wenn sie über unterschiedliche Modalitäten aufgenommen werden, schwer parallel verarbeitet werden können, wenn sie in beiden Fällen verbal oder räumlich kodiert sind. Als Beispiel fällt es schwer einen Text zu lesen (visuell, verbal) und parallel die Nachrichten im Radio mitzuverfolgen (auditiv, verbal).

Zudem unterscheidet Wickens (1984, 2002) auf Ausführungsebene zwischen einer *manuellen* und einer *sprachlichen* (vokalen) Ausführung. Da sich in experimentellen Untersuchungen zeigte, dass manuelle und sprachliche Handlungen kaum interferieren (u.a. McLeod, 1977; Wickens, 1980; Wickens & Liu, 1988) nimmt Wickens (2002) an, dass sie jeweils unterschiedliche Ressourcen beanspruchen (manuelle Handlungen räumliche Ressourcen und sprachliche Handlungen verbale Ressourcen).

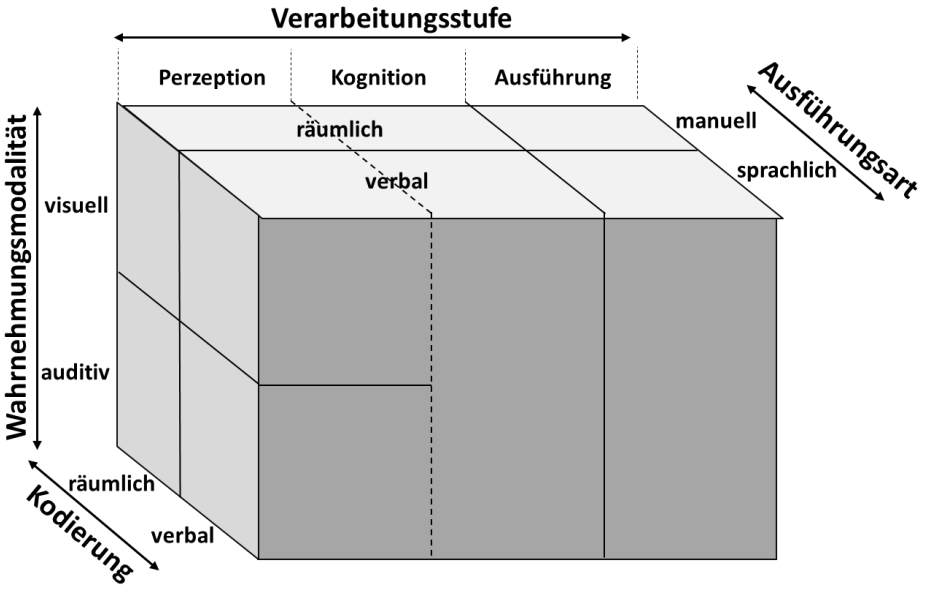


Abbildung 4. Modell zur Multiple Resource Theory nach Wickens (1984)

Für die Diagnose des Nutzerzustandes bedeutet dies, dass das technische System Informationen erhalten sollte, welche Aufgaben bearbeitet werden und welche Ressourcen diese beanspruchen. Aus diesen Informationen können nach Wickens (2002) die zu erwartenden Interferenzen und Leistungseinbußen bei Parallelaufgaben berechnet werden. Eine Bestimmung ist auch über Softwaretools möglich z.B. über W/INDEX (North & Riley, 1988), das in das Modellierungs-

werkzeug IPME eingegliedert ist (vgl. Abschnitt 2.4.3). In adaptiven Mensch-Maschine-Systemen können diese Daten genutzt werden, um zum Beispiel in Abhängigkeit der jeweiligen Ressourcenauslastung Informationen visuell oder auditiv, räumlich oder verbal darzustellen und so eine einseitige Auslastung kognitiver Ressourcen zu vermeiden (vgl. Fuchs et al., 2006).

Cognitive Task Load Modell

Das Cognitive Task Load (CTL-)Modell von Neerincx (2003) stellt dar, durch welche Faktoren die mentale Belastung systematisch variiert werden kann. Daher wird es in der vorliegenden Arbeit für die Modulierung der mentalen Beanspruchung in den durchgeführten experimentellen Untersuchungen zugrunde gelegt. Das CTL-Modell unterscheidet hinsichtlich der Aufgabenanforderungen zwischen drei Dimensionen, die sich auf die Belastung und somit auf die mentale Beanspruchung auswirken können: *Beschäftigte Zeit* („time occupied“), *Niveau der Informationsverarbeitung* („level of information processing“) und *Aufgabenwechsel* („task-set switches“). Abbildung 5 zeigt eine räumliche Anordnung der drei Dimensionen. Hierbei entspricht der grün markierte Bereich einer adäquaten Belastung, der rote Bereich einer zu hohen Belastung und der gelbe Bereich einer zu niedrigen Belastung.

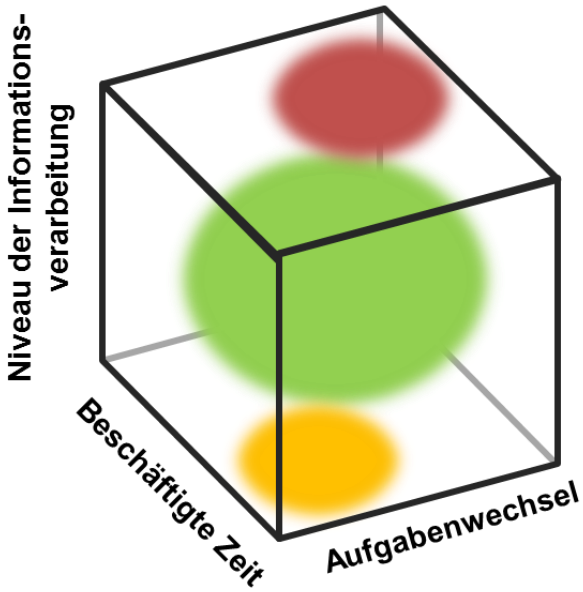


Abbildung 5. Die drei Dimensionen des Cognitive Task Load-Modells von Neerincx (2003) basierend auf einer Darstellung in DeGreef et al. (2009)

Die erstgenannte Dimension bezieht sich auf die Zeit, die für die Aufgabenbearbeitung benötigt wird. Eine hohe Ausprägung liegt dann vor, wenn der Operateur mit maximaler Geschwindigkeit arbeiten muss, um die Aufgaben in der verfügbaren Zeit zu erledigen. Dieser Faktor kann durch das Volumen der zu bearbeitenden Aufgaben variiert werden (DeGreef & Arciszewski, 2009).

Das Niveau der Informationsverarbeitung wird durch die Komplexität der Situation bestimmt. Hierbei wird unterschieden zwischen automatisierter Verarbeitung (geringe Belastung), Routineverfahren (mittlere Belastung) und Problemlösung/Handlungsplanung für neue Situationen (hohe Belastung). Diese Unterscheidung ist vergleichbar mit den Verhaltensebenen *Skills*, *Rules* und *Knowledge* nach Rasmussen (1983).

Die Dimension Aufgabenwechsel bezieht sich schließlich auf die Häufigkeit von Aufgabenwechseln, welche Verlagerungen der Aufmerksamkeit und eine Aktivierung unterschiedlicher Wissensinhalte erfordern.

Auswirkungen mentaler Beanspruchung

Mentale Beanspruchung kann sich sowohl positiv als auch negativ auf die Leistung eines Menschen auswirken (vgl. DIN EN 10 075-1, 2000). Im Sinne einer Aufwärmung oder Aktivierung kann sie anregend wirken. Andererseits können aber auch Ermüdung sowie ermüdungsähnliche Zustände (Monotonie, herabgesetzte Wachsamkeit, Sättigung; vgl. Abschnitt 3.1.4) und Stress (vgl. Abschnitt 3.1.2) resultieren. Um den Zusammenhang zwischen Beanspruchung und Leistung zu beschreiben, wird häufig das *Yerkes-Dodson-Law* (1908) herangezogen. Dieses Gesetz besagt, dass zwischen physischer Aktiviertheit (Arousal) und der Leistungsfähigkeit ein umgekehrt U-förmiger Zusammenhang besteht. Auf die mentale Beanspruchung bezogen, ergibt sich das in Abbildung 6 dargestellte Modell (nach Veltman & Jansen, 2006). Demzufolge ist die Leistung dann am höchsten, wenn die Beanspruchung in einem mittleren Bereich liegt. Bei sehr hoher und sehr geringer Beanspruchung (respektive Aktivierung) kann es hingegen aufgrund von Über- bzw. Unterforderung zu erheblichen Leistungsbeeinträchtigungen kommen. Durch das Investieren von Anstrengung ist es allerdings möglich, einem Abfallen der Leistung entgegenzuwirken (vgl. hierzu die Ausführungen zur Motivation in Abschnitt 3.1.3).

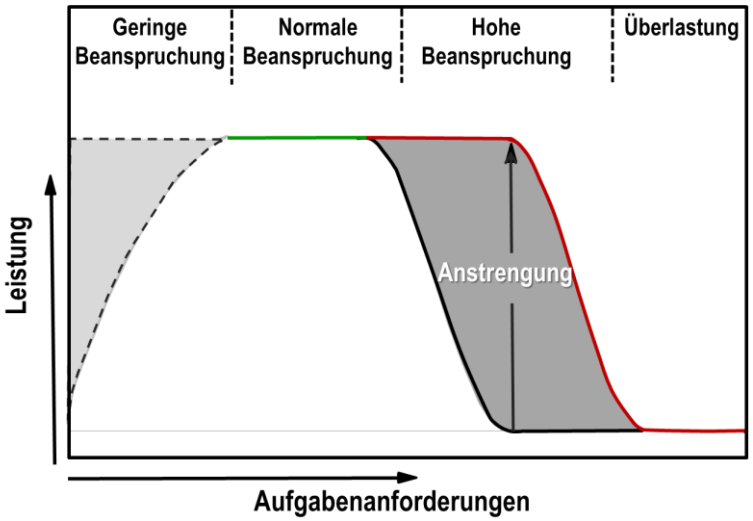


Abbildung 6. Darstellung des Zusammenhangs zwischen Aufgabenanforderungen und der Leistung nach Veltman & Jansen (2006)

Implikationen für die vorliegende Arbeit

Tabelle 9 fasst die wesentlichen Erkenntnisse zusammen, die aus den Beanspruchungsmodellen für die Berücksichtigung der mentalen Beanspruchung in adaptiven Systemen abgeleitet werden können. Geordnet sind diese in Hinblick auf relevante Einflussfaktoren, kritische Ausprägungen, die sich negativ auf die Leistungsfähigkeit auswirken können, weitere wichtige Unterscheidungen oder Ausprägungsformen und Auswirkungen auf andere Zustandsdimensionen.

Um die Relevanz für die adaptive Systemgestaltung aufzuzeigen, wird zuletzt auf Möglichkeiten hingewiesen, den Nutzer auf dieser Basis adaptiv zu unterstützen. Dieser Aspekt wird im Promotionsvorhaben von Fuchs noch eingehender beleuchtet (vgl. Abschnitt 1.4).

Tabelle 9. Erkenntnisse aus den Theorien und Modellen zur mentalen Beanspruchung

Betrachtete Aspekte	Ergebnisse der Analyse
Relevante Einflussfaktoren	• Belastungsfaktoren: Aufgabe, Kontext, Umgebungsbedingungen, Systemeigenschaften • Individuelle Faktoren: z.B. Kenntnisse, Fähigkeiten, Erfahrung
Kritische Ausprägungen	• Hohe und geringe mentale Beanspruchung
Wichtige Unterscheidungen	• Perzeptive, kognitive, psychomotorische Beanspruchung • Ressourcenauslastung räumlich vs. verbal
Primäre Auswirkungen	• Müdigkeit • Aufmerksamkeit (i.S.v. herabgesetzter Wachsamkeit) • Stress
Mögliche Unterstützung durch das adaptive System	• Entlastung bei zu hoher Beanspruchung (beschrieben in Fuchs & Schwarz, 2017; vgl. auch Ansätze zur Adaptiven Automation in Abschnitt 2.1.2) • Veränderung der Stimuluspräsentation bei Überlastung eines Sinneskanals

3.1.2 Emotionaler Zustand

Im Folgenden wird das Konstrukt des emotionalen Zustands näher betrachtet, wozu in der vorliegenden Arbeit, dem Vorschlag von Lazarus (1993) folgend, auch das Konstrukt „Stress“ gezählt wird.

Definition und Klassifikation

Emotionen können als aktuelle psychische Zustände beschrieben werden, die durch die bewusste und/oder unbewusste Wahrnehmung eines Objekts oder einer Situation ausgelöst werden und mit *physiologischen* Veränderungen, spezifischen *Kognitionen*, subjektivem *Gefühlserleben* und einer Veränderung der *Verhaltensbereitschaft* einhergehen (vgl. Kleiniginna & Kleiniginna, 1981a).

Bei dem Konzept des emotionalen Zustands handelt es sich um ein multidimensionales Konstrukt, das sich in verschiedene spezifische Emotionen untergliedert. In kategorialen Beschreibungsansätzen werden verschiedene Basisemotionen unterschieden, wie Freude, Traurigkeit, Furcht, Ärger, Überraschung, Ekel (vgl. Ekman, Friesen, & Ellsworth, 1982). Bei dem Circumplex-Modell von Russell (1980) handelt es sich um einen dimensionalen Beschreibungsansatz. Demzufolge können Emotionen auf Basis der Dimensionen *Valenz* mit den Polen angenehm und unangenehm und der *Erregung* (engl. Arousal) mit den Polen erregt und schläfrig klassifiziert werden. Das Modell ist in Abbildung 7 mit den äquivalenten deutschen Begriffen veranschaulicht. Russell (1980) konnte empirisch belegen, dass sich die meisten der von ihm untersuchten Emotionen in einem durch diese beiden Dimensionen aufgespannten Koordinatensystem anordnen lassen. Da einige Emotionen jedoch ähnliche Ausprägungen in Bezug auf Valenz und Arousal aufweisen (z.B. Ärger und Angst als Emotionen mit negativer Valenz und hohem Arousal), wird zur weiteren Differenzierung die Dimension „Annäherung/Vermeidung“ vorgeschlagen (Elliott, Eder, &

Harmon-Jones, 2013). Sie bezieht sich auf die Verhaltens-bereitschaft, die mit Emotionen verbunden ist. So ist Ärger mit einer Tendenz zur Annäherung und Angst mit einer Tendenz zur Vermeidung verbunden (Mauss & Robinson, 2009, Carver & Harmon-Jones, 2009). Valenz und Arousal können über physiologische und verhaltensbasierte Maße erfasst werden. Wie sich in Abschnitt 2.3.3 zeigte, kann die Valenz z.B. durch die Gesichtsmuskelaktivität und der Erregungszustand z.B. durch die Hautleitfähigkeit und die Herzaktivität erfasst werden.

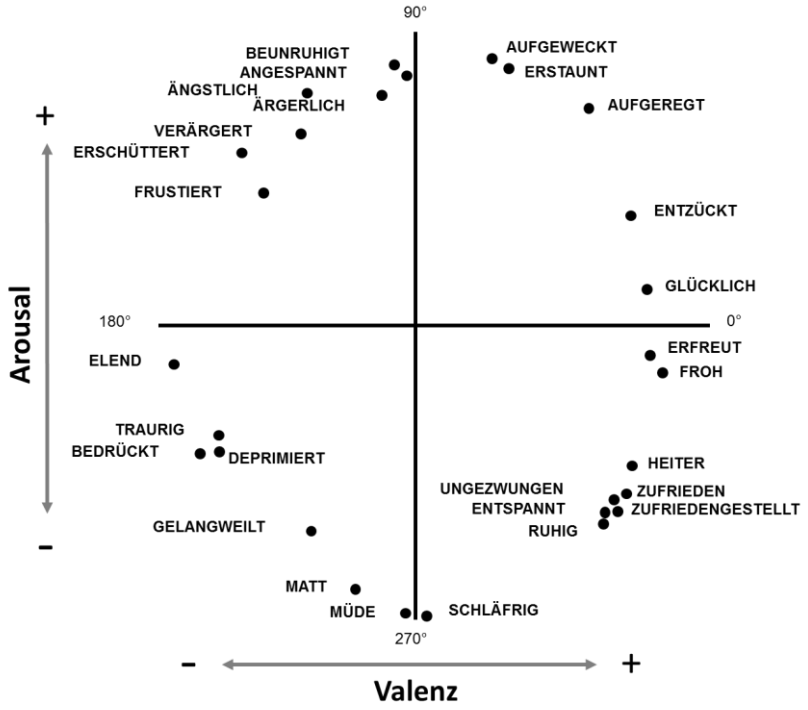


Abbildung 7. Circumplex-Modell basierend auf Russell, 1980

Einflussfaktoren und Wirkzusammenhänge

Eine Vielzahl an Theorien beschäftigt sich mit den Wirkzusammenhängen zwischen emotionsauslösenden Reizen, Veränderungen in Physiologie und Kognition und dem Emotionserleben (u.a. James-Lange-Theorie – James, 1884; Cannon & Bard-Theorie – Cannon 1924, Theorie von Maranon, 1924, Zwei-Faktoren-Theorie von Schachter & Singer, 1962, Kognitive Bewertungstheorie von Lazarus, 1966).

Von besonderem Interesse für die vorliegende Arbeit sind Theorien und Modelle, die sich auf die Entstehung und die Auswirkungen von emotionalen Zuständen beziehen, welche die Leistung negativ beeinflussen. In diesem Kontext sind die Arbeiten von Lazarus und Kollegen im Bereich der psychologischen Stressforschung bedeutsam (z.B. Lazarus & Eriksen, 1952, Lazarus 1966). Nach Lazarus (1966) ist Stress das Resultat von Stressoren, die auf den Menschen einwirken, wobei der gleiche Stressor bei unterschiedlichen Personen aufgrund von individuellen Faktoren (z.B. Bewältigungsstrategien, Persönlichkeitseigenschaften) unterschiedliche Stressreaktionen hervorrufen kann. Er unterscheidet dabei zwischen drei Arten von Stress: *Schädigung* („harm“), *Bedrohung* („threat“) und *Herausforderung* („challenge“). Schädigung und Bedrohung sind üblicherweise mit negativen Auswirkungen auf die Leistung verbunden (vgl. „Distress“, Selye, 1974) während Herausforderung leistungsförderliche Effekte hat (vgl. „Eustress“, Selye, 1974).

In einer Pfadanalyse erstellten und überprüften Nicholls et al. (2012) basierend auf den Erkenntnissen von Lazarus (1999) ein Modell, das beschreibt, welche Faktoren Einfluss auf die Entstehung von Emotionen und Stress im Sport haben, und wie diese mit der Leistung zusammenhängen (vgl. Abbildung 8).

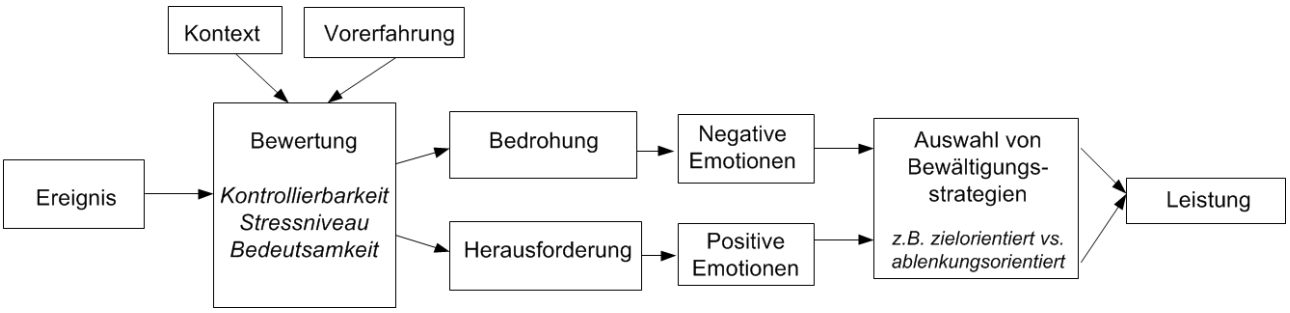


Abbildung 8. Vereinfachte Darstellung des Pfadmodells von Nicholls et al. (2012)

Dem Modell zufolge entstehen Emotionen durch eine Bewertung von Ereignissen hinsichtlich ihrer *Kontrollierbarkeit*, des *Stressniveaus* und ihrer *Bedeutsamkeit*. Sowohl der Kontext als auch Vorerfahrungen fließen in die Bewertung mit ein. Auf Basis dieser Bewertung wird das Ereignis entweder als Bedrohung oder als Herausforderung wahrgenommen. Nicholls et al. (2012) stellten fest, dass Unkontrollierbarkeit, ein hohes Stressniveau und eine hohe Bedeutsamkeit des Ereignisses die Wahrnehmung eines Ereignisses als Bedrohung begünstigen. Ein solches bedrohliches Ereignis kann in Mensch-Maschine-Systemen zum Beispiel das Eintreten einer unvorhergesehenen Notsituation sein (Wickens, 1996a). Bedrohliche Ereignisse begünstigen dem Modell zufolge wiederum die Entstehung negativer Emotionen, welche in der Folge die Wahrscheinlichkeit erhöhen, dass Bewältigungsstrategien angewandt werden, die für eine erfolgreiche Aufgabenbewältigung hinderlich sind (z.B. ablenkungsorientierte Bewältigungs-strategien). Demgegenüber fördern Ereignisse, die als Herausforderung wahrgenommen werden, das Auftreten positiver Emotionen. Diese erhöhen die Wahrscheinlichkeit, dass Bewältigungs-strategien angewandt werden, die für die Leistung förderlich sind, z.B. der Einsatz ziel- bzw. aufgabenorientierter Bewältigungsstrategien.

Auswirkungen von negativen emotionalen Zuständen und Stress

In verschiedenen empirischen Studien wurden die Auswirkungen von negativen emotionalen Zuständen sowie von Stress auf die Prozesse der Informationsverarbeitung und die Leistung untersucht. Ein differenzierter Überblick findet sich in Staal (2004) und Wickens (1996a). Es zeigte sich, dass Stress, insbesondere in Verbindung mit Angst, Aufmerksamkeits- und Gedächtnis-prozesse beeinflusst und somit die Leistung beeinträchtigt. In diesem Zusammenhang wurde der Effekt beobachtet, dass Stress die Wahrnehmung und Aufmerksamkeit einengt, so dass ein so genannter Tunnelblick entsteht. Dies kann sich so auswirken, dass Signale im peripheren Blickfeld seltener wahrgenommen werden oder dass sich die Aufmerksamkeit ausschließlich auf die Stimuli fokussiert, die (subjektiv) am wichtigsten zu sein scheinen (Wickens, 1996a). Dies muss sich nicht zwingend negativ auf die Leistung auswirken. Probleme ergeben sich jedoch dann, wenn es die Aufgabe erfordert, dass verschiedene Informationen berücksichtigt und integriert werden müssen,

oder wenn die subjektive Priorisierung fehlerhaft ist (Staal, 2004, Wickens, 1996a; vgl. auch die Ausführungen zur Aufmerksamkeit in Abschnitt 3.1.5).

Negative Auswirkungen auf Gedächtnisprozesse wurden z.B. von Ashcraft & Kirk, 2001 und Eysenck (1997) festgestellt. Diese Effekte können dadurch erklärt werden, dass negative Emotionen kognitive Ressourcen beanspruchen, die in der Folge nicht mehr für die Aufgabenbearbeitung zur Verfügung stehen. Einerseits aktivieren ängstliche Gedanken weitere ängstliche Gedanken, was zu Rumination führt und die Informationsverarbeitung blockiert (Bower, 1981). Andererseits werden durch die Anwendung von Bewältigungsstrategien zusätzlich kognitive Kapazitäten verbraucht (Richards & Gross, 2000). Dies kann dazu führen, dass Informationen weitgehend nur auf Basis von Heuristiken bewertet werden und Informationen, die den kognitiven Erwartungen und Hypothesen widersprechen, ignoriert werden („confirmation bias“).

Aber auch negative emotionale Zustände, die mit einem geringeren Erregungszustand verbunden sind als Angst, können die Leistung negativ beeinträchtigen. Nach Waterhouse & Child (1953) kann Frustration zu Verhalten führen, das mit der Aufgabenbearbeitung interferiert und somit die Qualität der Leistung beeinträchtigt. Ein anderer Problemzustand insbesondere bei hochautomatisierten Überwachungs- und Kontrollaufgaben ist Langeweile. So gibt es Belege, dass Langeweile Stress und Frustration auslösen kann (Scerbo, 2001) und sich negativ auf die Motivation, Müdigkeit, und Aufmerksamkeit auswirkt, was insbesondere in sicherheitskritischen Bereichen mit gravierenden Konsequenzen verbunden sein kann (vgl. Cummings, Gao, & Thornburg, 2016).

Implikationen für die vorliegende Arbeit

In Tabelle 10 sind die wesentlichen Erkenntnisse zusammengefasst, die aus den betrachteten Theorien und Modellen zum emotionalen Zustand für die Nutzerzustandsdiagnose und die adaptive Systemgestaltung gewonnen wurden (in analoger Struktur zu Tabelle 9).

Tabelle 10. Erkenntnisse aus den Theorien und Modellen zum emotionalen Zustand

Betrachtete Aspekte	Ergebnisse der Analyse
Relevante Einflussfaktoren	• Bedeutsame Ereignisse • Individuelle Faktoren (z.B. Bewältigungsstrategien, Persönlichkeitseigenschaften)
Kritische Ausprägungen	• Negative Emotionen (z.B. Angst, Frustration, Langeweile) • Stress i.S.v. Distress
Wichtige Unterscheidungen	• Valenz (positiv-negativ) • Arousal/Erregungszustand (hoch-gering) • (Annäherung/Vermeidung)
Primäre Auswirkungen	• Aufmerksamkeit • Informationsverarbeitung • Handlungsbereitschaft (Motivation)
Mögliche Unterstützung durch das adaptive System	• Unterstützung bei Bewältigung stressauslösender Ereignisse geben (z.B. Handlungsmöglichkeiten aufzeigen)

3.1.3 Motivation

Zur Motivation existieren vielfältige Theorien und Modelle, wobei sich diese teilweise mit sehr spezifischen Aspekten, wie der Motivation im Lern- oder im Arbeitskontext beschäftigen. Im Folgenden werden die Theorien und Modelle näher dargestellt, die den höchsten Erkenntnisgewinn für die adaptive Systemgestaltung versprechen. Hierbei ist insbesondere der Bezug der Motivation zur Anstrengungsregulierung von Bedeutung.

Definition und Klassifikation

Motivation bezieht sich auf psychologische Prozesse, die für die Initiierung, Ausrichtung und Aufrechterhaltung von zielgerichteten Verhaltensweisen zuständig sind (Greenberg & Baron, 2008; Buchanan & Huczynski, 1997). Nach Zimbardo (1980, zitiert in Kleiniginna & Kleiniginna, 1981b) ist Motivation durch vier Merkmale charakterisiert: (a) Energie, Arousal, (b) Ausrichtung der Anstrengung auf ein bestimmtes Ziel, (c) selektive Aufmerksamkeit für relevante Stimuli, (d) Organisation des Verhaltens in Form von Verhaltensmustern oder -sequenzen, (e) Beibehaltung des Verhaltens bis sich die auslösenden Bedingungen geändert haben.

Eine grundlegende Frage ist, was den Menschen zu diesem zielgerichteten Verhalten motiviert. Erklärbar ist dies durch extrinsische und intrinsische Anreize. Bei der *extrinsischen Motivation* wird die Motivation durch externe Faktoren gesteuert, z.B. durch die Aussicht auf Belohnungen oder durch das Vermeiden von Strafe. Demgegenüber ist *intrinsische Motivation* dadurch charakterisiert, dass Handlungen um ihrer selbst willen (z.B. weil es Spaß macht oder eine Herausforderung darstellt) getätigt werden. Nach Vallerand (1997) ist intrinsische Motivation mit positiven Emotionen verbunden während extrinsische Motivation zu Anspannung und Druck führen kann. Den Gegenpol zu diesen beiden Motivationsarten bildet die *Amotivation*, die sich auf das Fehlen von Motivation bezieht. Handlungen werden hierbei ohne Intentionen und nicht zielgerichtet ausgeführt (Deci & Ryan, 1985, Vallerand, 1997).

Erwartung-Wert-Theorien

In den Erwartung-Wert-Theorien (z.B. Vroom, 1964; Porter & Lawler, 1968) wird davon ausgegangen, dass sich die Motivation für eine bestimmte Handlung oder die Erbringung einer Leistung nicht nur aus den erwarteten intrinsischen oder extrinsischen Belohnungen sondern aus dem Produkt von *Erwartung* und *Wert* zusammensetzt. Erwartung entspricht dabei der subjektiv eingeschätzten Wahrscheinlichkeit, durch die Handlung ein Ziel zu erreichen und der Wert bezieht sich auf den Wert, der dem Ziel bzw. den erwarteten intrinsischen und extrinsischen Belohnungen beigemessen wird. Unter mehreren Handlungsalternativen wird demzufolge die ausgewählt, die den höchsten Erwartungswert besitzt.

Das Modell von Porter und Lawler (1968) ist in Abbildung 9 dargestellt. In dem Modell bezieht sich die Motivation auf die Bereitschaft, sich anzustrengen, um eine Leistung zu erbringen. Dabei ist die Anstrengungsbereitschaft umso höher, je höher Erfolgswahrscheinlichkeit und Wert der Belohnung einer Handlung eingeschätzt werden. Der Einfluss der Anstrengung auf die Leistung wird zusätzlich durch die Fähigkeit und die Rollenwahrnehmung moderiert. Aus der Leistung ergeben sich daraufhin bestimmte intrinsische und extrinsische Belohnungen, welche die

Zufriedenheit steigern. Es wird angenommen, dass die Zufriedenheit in einer Feedback-Schleife den künftig angenommenen Wert der Belohnungen bestimmt, während die Leistung Einfluss auf die Einschätzung der Erfolgswahrscheinlichkeit nimmt.

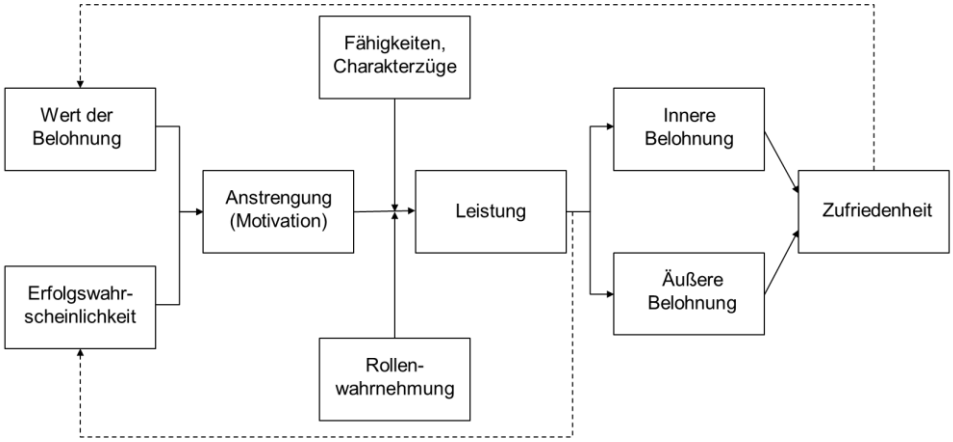


Abbildung 9. Modell von Porter & Lawler (1968) nach Pelz (2004)

Zielsetzungstheorien

Die Zielsetzungstheorien beschäftigen sich damit, dass Motivation und Arbeitsleistung wesentlich durch das Festlegen von Zielen bestimmt werden. Locke & Latham (1990) wiesen anhand verschiedener empirischer Untersuchungen nach, dass Ziele insbesondere dann eine motivierende Wirkung haben, wenn sie herausfordernd und präzise gestaltet sind und wenn regelmäßig Rückmeldungen über Zielfortschritte gegeben werden. Dass sich Ziele motivationsfördernd auswirken, lässt sich damit erklären, dass Ziele – wie in einem Regelkreissystem – einen Spannungszustand erzeugen, der die Bereitschaft (Motivation) sich anzustrengen erhöht (siehe nachfolgende Ausführungen zu „Auswirkungen von Motivation“).

Auswirkungen von Motivation

Die Zusammenhänge zwischen Motivation, Anstrengung und Leistung können mit Hilfe der zuvor erwähnten regelkreisbasierten Betrachtung erklärt werden. Veltman & Jansen (2006) beschreiben den Prozess so, dass der menschliche Organismus Ist-Soll-Vergleiche vornimmt, in denen die gegenwärtige Leistung mit dem Zielzustand, also der für die Zielerreichung notwendigen Leistung, verglichen wird. Eine Diskrepanz zwischen Ist und Soll motiviert (im Normalfall) dazu, mehr Anstrengung zu investieren, was die Intensität der Informationsverarbeitung und damit die Leistung erhöht (vgl. Abbildung 6 in Abschnitt 3.1.1). Brehm & Self (1989) sowie Wright (1996) stellten dazu fest, dass der Grad der Anstrengung, der investiert wird, linear mit der Schwierigkeit der Zielerreichung, (Aufgabenschwierigkeit) steigt. Dieser Zusammenhang kann durch die Erwartungs-Wert-Theorien nicht erklärt werden, da eine höhere Schwierigkeit weder die Erfolgserwartung noch den Wert der Belohnung erhöht. Die Autoren gehen daher davon aus, dass die in den Erwartung-Wert-Theorien postulierten Einflussfaktoren die potentielle Motivation determinieren und damit eine Obergrenze der Anstrengungsbereitschaft festlegen, wobei sich die tatsächliche Anstrengung nach der Aufgabenschwierigkeit richtet und mit zunehmender Schwierigkeit bis zu dieser Obergrenze zunimmt. Nach dem Überschreiten der Obergrenze geht die Anstrengung zurück und

bleibt auf einem geringen Niveau. Dies kann auch dann der Fall sein, wenn das Ziel als nicht erreichbar bzw. die Aufgabe als unlösbar eingeschätzt wird (vgl. Abfallen der Leistungskurve in Abbildung 6, Abschnitt 3.1.1). Ebenso weisen Cropanzano, James, & Citera (1993) darauf hin, dass Zielanforderungen, die die eingeschätzten Fähigkeiten übersteigen, einen amotivationalen Zustand hervorrufen. Dies führt dazu, dass weniger Anstrengung investiert wird und geht mit negativen Emotionen einher, die sich, wie in Abschnitt 3.1.2 näher ausgeführt wurde, ebenfalls negativ auf die Leistung auswirken können.

Implikationen für die vorliegende Arbeit

Wie sich zeigt, ist Motivation eng mit der Anstrengungsregulierung verknüpft. Durch die theoretische Analyse bestätigt sich die Erkenntnis aus Abschnitt 2.4.4, dass adaptive Systeme (erst) dann Unterstützung bieten sollten, wenn die Anstrengung des Operateurs nicht ausreicht, die Aufgaben und Ziele zu erreichen. Weitere Ergebnisse der theoretischen Analyse sind in Tabelle 11 aufgeführt.

Tabelle 11. Erkenntnisse aus den Theorien und Modellen zur Motivation

Betrachtete Aspekte	Ergebnisse der Analyse
Relevante Einflussfaktoren	• Ziele • Individuelle Faktoren (z.B. Fähigkeiten) • Erwartete Konsequenzen der Handlungen (z.B. Belohnungen) • Aufgabenschwierigkeit
Kritische Ausprägungen	• Geringe Motivation/Amotivation
Wichtige Unterscheidungen	• Intrinsische und extrinsische Motivation
Primäre Auswirkungen	• Anstrengung • Emotionaler Zustand • Aufmerksamkeit
Mögliche Unterstützung durch das adaptive System	• Unterstützung bei der Zielerreichung • Rückmeldung über Fortschritt bei Zielerreichung

3.1.4 Müdigkeit

Mit dem Konstrukt Müdigkeit beschäftigen sich, wie Hockey (2013) herausstellt, in den letzten Jahrzehnten nur wenige psychologische Theorien. Mehr Beachtung kommt im medizinischen Bereich der chronischen Müdigkeit zu, die jedoch nicht im Fokus der vorliegenden Arbeit steht. Dennoch ermöglicht die theoretische Analyse, einige wichtige Erkenntnisse zu wesentlichen Einflussfaktoren und Auswirkungen von Müdigkeit zu gewinnen, die für die adaptive Systemgestaltung von Bedeutung sind.

Definition und Klassifikation

Zu Müdigkeit existieren nach Phillips (2015) unterschiedliche Definitionen und Operationalisierungen, die sich häufig nur auf bestimmte Aspekte, wie die Physiologie, das Erleben oder die Leistung beziehen. In einem Versuch, die verschiedenen Auffassungen zu integrieren, schlägt Phillips (2015) vor, Müdigkeit als ein multidimensionales Konzept zu betrachten, das einen

suboptimalen psychophysiologischen Zustand charakterisiert. Dieser hängt ab von der Form, Dynamik und dem Kontext (z.B. individuellen Faktoren, Umweltbedingungen, Schlafhistorie, circadianen Effekten) körperlicher und psychischer Anstrengungen. Er führt zu Veränderungen in Strategien oder Ressourcenverwendungen und wirkt sich so auf die Leistung aus.

Je nach Art der Anstrengungen kann zwischen Müdigkeit, die durch mentale Anstrengung hervorgerufen wird, und Müdigkeit, die durch physische Anstrengung bedingt ist, unterschieden werden (Gawron, French, & Funke, 2001; Beaumont et al., 2004), wobei beide Formen auch gemeinsam auftreten können (Phillips, 2015). Ein mit der Müdigkeit verwandtes Konzept ist die *Schläfrigkeit* (engl. sleepiness), die durch Schlafmangel hervorgerufen wird und das Bedürfnis nach Schlaf bezeichnet (vgl. Thorpy & Billiard, 2011). May & Baldwin (2009) integrieren in ihrem Modell, das sie für den Bereich der Fahrzeugführung entwickelten, diese Konzepte. Das Modell ist in modfizierter Form – abstrahiert vom Kontext der Fahrzeugführung – in Abbildung 10 wiedergegeben. In Abhängigkeit von der Art der auslösenden Bedingungen unterscheidet das Modell zwischen *schlafbezogener Müdigkeit* (sleep-related fatigue) und *aufgabenbezogener Müdigkeit* (task-related fatigue), wobei die aufgabenbezogene Müdigkeit in Übereinstimmung mit Desmond & Hancock (2001) und Gimeno, Cerezuela, & Montanes (2006) noch weiter in eine aktive und eine passive Form unterteilt wird. Während *aktive aufgabenbezogene Müdigkeit* durch hohe Beanspruchung hervorgerufen wird, entsteht die *passive aufgabenbezogene Müdigkeit* bei stark monotonen Aufgaben mit sehr geringer Beanspruchung und Vorhersagbarkeit. Die *schlafbezogene Müdigkeit* resultiert demgegenüber aus der vorangegangenen Schlafdauer und -qualität und aus dem circadianen Rhythmus, der die Leistungsfähigkeit in Abhängigkeit der Tageszeit moduliert. Sie kann sich dem Modell zufolge durch aufgabenbezogene Müdigkeit weiter verschlechtern.

Eine Reduzierung schlafbezogener Müdigkeit ist in erster Linie durch Schlaf, ggf. auch durch Stimulantien wie Koffein möglich (Philip et al., 2006; Reyner & Horne, 2000). Der passiven und aktiven Form aufgabenbezogener Müdigkeit kann jedoch bereits durch Änderungen der externen Bedingungen (z.B. Automationsgrad verändern, Anregungen schaffen) entgegengewirkt werden (May & Baldwin, 2009). Dies lässt die Schlussfolgerung zu, dass insbesondere aufgabenbezogene Müdigkeit durch adaptive Systemgestaltung adressiert werden kann.

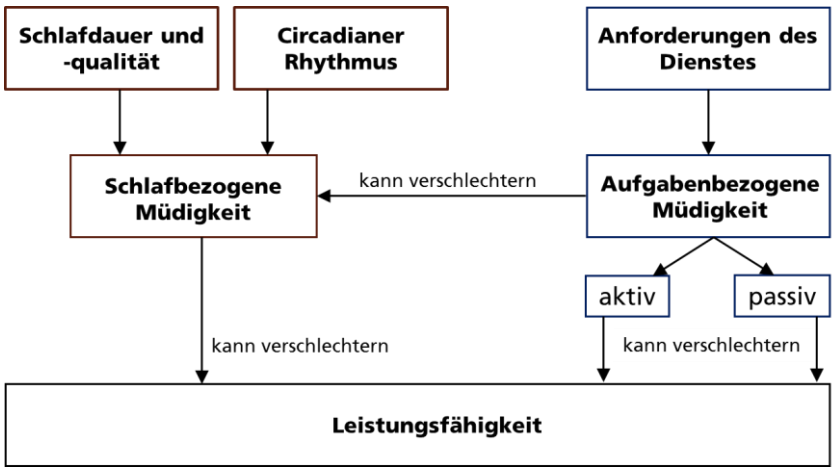


Abbildung 10. Differenzierung zwischen schlafbezogener und aufgabenbezogener Müdigkeit nach May & Baldwin (2009)

Auswirkungen von Müdigkeit

Die Auswirkungen von Müdigkeit auf die kognitive Leistungsfähigkeit wurden vielfach auf Basis von Schlafentzug untersucht und beziehen sich daher vorwiegend auf schlafbezogene Müdigkeit (vgl. Review mehrerer Studien in Miller, Matsangas, & Shattuck, 2008; Krueger, 1989). Die Studien unterscheiden sich in der Länge des Schlafentzugs und der Art der Aufgaben, die zur Leistungsmessung herangezogen wurden. Überwiegend konnten die Studien deutliche Einbußen hinsichtlich der kognitiven Fähigkeiten belegen. Diese betreffen das logische Denken und die Entscheidungsfindung, Vigilanz und Aufmerksamkeit, das Situationsbewusstsein und die Reaktionszeit (vgl. auch Hursh, Balkin, Miller, & Eddy, 2004).

Aufgabenbezogene Müdigkeit wirkt sich ebenfalls negativ auf die Leistung aus, wenn auch schwächer als in Kombination mit schlafbezogener Müdigkeit (vgl. Philip et al., 2005). In Fahrsimulatorexperimenten zeigte sich, dass inbesondere bei langandauernden monotonen Fahrten passive aufgabenbezogene Müdigkeit zu Einbußen in Vigilanz und Fahrverhalten führt (Rossi, Gastaldi, & Gecchele, 2011; Thiffault & Bergeron, 2003). Nach Gimeno et al. (2006) bezieht sich die passive Form der Müdigkeit auf einen Zustand reduzierter Aufmerksamkeit, der mit einem Absinken physiologischer Aktivität (geringes Arousal), einem Gefühl der Trägheit und reduziertem Situationsbewusstsein einhergeht. Verwandte Konzepte sind *Hypovigilanz* (vgl. Larue, Michael, & Rakotonirainy, 2011, sowie Abschnitt 3.1.5) und *Langeweile* (vgl. Hockey, 2013 und Abschnitt 3.1.2). Aktive aufgabenbezogene Müdigkeit entsteht demgegenüber als Folge von lange andauernder hoher Beanspruchung und geht somit mit ähnlichen Auswirkungen wie diese einher (vgl. May & Baldwin, 2009, Gimeno et al., 2006 sowie Abschnitt 3.1.1).

Implikationen für die vorliegende Arbeit

Tabelle 12 fasst die Ergebnisse zusammen, die aus der theoretischen Analyse für die Berücksichtigung von Müdigkeit in adaptiven Systemen abgeleitet werden können.

Tabelle 12. Erkenntnisse aus den Theorien und Modellen zur Müdigkeit

Betrachtete Aspekte	Ergebnisse der Analyse
Relevante Einflussfaktoren	• Aufgabenart und –dauer (aufgabenbezogene Müdigkeit) • Schlafmenge/-qualität, Tageszeit (schlafbezogene Müdigkeit)
Kritische Ausprägungen	• Hohe Müdigkeit

Wichtige Unterscheidungen	• Aufgabenbezogene Müdigkeit
	  • Aktiv
	  • Passiv
	• Schlafbezogene Müdigkeit

Primäre Auswirkungen	• Aufmerksamkeit • Situationsbewusstsein • Beanspruchung
Mögliche Unterstützung durch das adaptive System	• Erholungsmöglichkeiten anbieten bei aktiver aufgabenbezogener Müdigkeit • Anregung schaffen bei passiver aufgabenbezogener Müdigkeit (vgl. Fuchs & Schwarz, 2017)

3.1.5 Aufmerksamkeit

Bei der Aufmerksamkeit können zwei Ausprägungsformen unterschieden werden, die bei der Mensch-Maschine-Interaktion Problemzustände auslösen können: Die (momentane) Aufmerksamkeit, die auf bestimmte Reize gerichtet ist und die Daueraufmerksamkeit (Vigilanz), die eine überdauernde Aufnahmebereitschaft von Informationen/Reizen charakterisiert. Zunächst wird die momentane gerichtete Aufmerksamkeit näher betrachtet. Anschließend wird auf die Vigilanz genauer eingegangen (S. 56).

Definition und Klassifikation

Aufmerksamkeit kann als ein psychischer Zustand beschrieben werden, der durch gesteigerte Wachheit und Aufnahmebereitschaft charakterisiert ist und bei dem sich das Bewusstsein auf bestimmte Informationen (Objekte, Prozesse, Gedanken) konzentriert (Schaub, 2008). Andere Informationen werden hingegen von der bewussten Wahrnehmung ausgeschlossen, so dass in diesem Zusammenhang auch der Begriff *selektive Aufmerksamkeit* verwendet wird (Johnston & Dark, 1986). Ressourcentheoretische Ansätze erklären diesen Vorgang damit, dass der menschliche Organismus nur begrenzte Ressourcen zur Informationsverarbeitung besitzt. Um einer Überlastung des kognitiven Systems vorzubeugen, werden einkommende Informationen in einem Filterprozess vorselektiert, bevor diese ins Kurzzeitgedächtnis gelangen und bewusst wahrnehmbar werden (vgl. Filtertheorie von Broadbent, 1958).

Einflussfaktoren und Wirkzusammenhänge

Die an Aufmerksamkeitsprozessen beteiligten Komponenten werden im SEEV-Modell von Wickens & McCarley (2008) näher beschrieben (Abbildung 11). *SEEV* stellt ein Akronym dar, das sich aus den Wörtern *Salience* (Salienz, Auffälligkeit), *Effort* (Anstrengung), *Expectancy* (Erwartung) und *Value* (Wert) zusammensetzt. Diese Konzepte werden als primäre Faktoren angesehen, die bewirken, dass sich die Aufmerksamkeit eines erfahrenen Operateurs selektiv bestimmten Informationsquellen zuwendet. Wickens & McCarley (2008) zufolge sind die Faktoren *Erwartung* und *Wert* einem mentalen Modell inhärent, das der Informationsverarbeitung zugrunde liegt. Sie steuern die Aufmerksamkeit somit in einem Top-Down-Prozess von innen. Die Erwartung bezieht sich auf die Wahrscheinlichkeit, die für das Vorhandensein oder Auftreten einer bestimmten Information angenommen wird. Der Wert bezieht sich auf die Nützlichkeit, die der Aufnahme einer bestimmten Information zugeschrieben wird, bzw. er resultiert daraus, welche (negativen) Konsequenzen für die Nichtaufnahme dieser Information angenommen werden.

Demgegenüber stellen die *Salienz*, also die Auffälligkeit der dargebotenen Informationen, und die für die Informationsaufnahme benötigte *Anstrengung* Faktoren dar, die extern durch die Umwelt gesteuert sind und somit durch Bottom-Up-Prozesse die Aufmerksamkeit beeinflussen. Die Anstrengung bei der Informationsaufnahme lässt sich dabei operationalisieren als die Zeit, die benötigt wird, um eine gewünschte Information zu finden. Es wird davon ausgegangen, dass die Salienz von Merkmalen positiven Einfluss hat, während die Anstrengung (z.B. aufgrund von Clutter auf dem Display) eine inhibitorische Wirkung besitzt. Hinzu kommen Anstrengungen, die für konkurrierende Aufgaben aufgebracht werden müssen und ebenfalls inhibitorisch die Aufmerksamkeit beeinflussen.

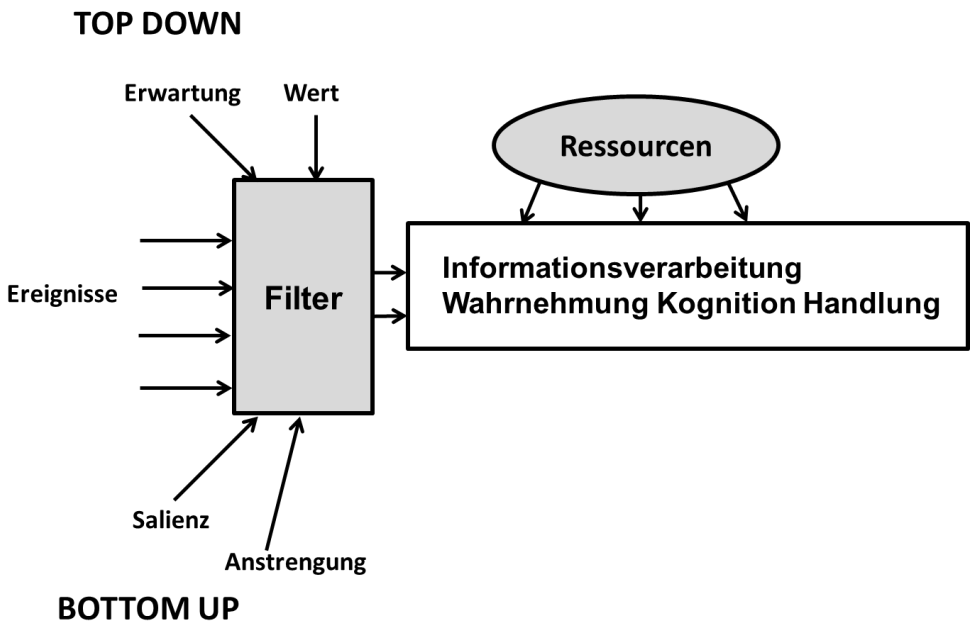


Abbildung 11. SEEV-Modell zur Aufmerksamkeit basierend auf Wickens & McCarley (2008)

Auswirkungen selektiver Aufmerksamkeit

Selektive Aufmerksamkeit kann zu Phänomenen führen, die als *inattentional blindness* (Mack & Rock, 1998) und *inattentional deafness* (MacDonald & Lavie, 2011; Molloy et al., 2015) bekannt sind. Hierbei wurde festgestellt, dass visuelle bzw. akkustische Stimuli trotz hoher Salienz nicht wahrgenommen werden, wenn diese unerwartet auftreten und die Aufmerksamkeit bereits durch eine Beobachtungsaufgabe gebunden ist. Ein eindrucksvolles Beispiel ist das „Gorilla-Experiment" von Simons & Chabris (1999).

Die Lenkung der Aufmerksamkeit im Top-Down-Prozess kann außerdem dazu führen, dass Informationen, welche die eigenen Erwartungen bestätigen, stärker beachtet werden als Informationen, die den Erwartungen widersprechen. Dieser Effekt wird als *confirmation bias* oder auch als *cognitive tunneling* bezeichnet (Wickens, Gordon & Liu, 2004). Problematisch ist diese hypothesenbasierte Aufmerksamkeitslenkung dann, wenn das zugrunde liegende mentale Modell oder die Erwartungen fehlerhaft sind und dadurch Alternativhypothesen unzureichend berücksichtigt werden.

Wie in Abschnitt 3.1.2 dargestellt wurde, können negative Emotionen, wie Angst oder Stress zu unerwünschten perzeptiven und kognitiven Einschränkungen des Aufmerksamkeitsfokus führen. Wichtige Informationen werden dabei nicht oder ungenügend wahrgenommen oder evaluiert. Das Konzept des *attentional tunneling* nach Wickens (2005) und Wickens & Alexander (2009) fasst die verschiedenen Aufmerksamkeitsprobleme, die durch Stress hervorgerufen werden können, wie folgt zusammen: *„We can offer a rough definition of attentional tunneling as the allocation of attention to a particular channel of information, diagnostic hypothesis or task goal, for a duration that is longer than optimal, given the expected cost of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks"* (Wickens, 2005, S.1).

Wickens (1996a) führt als Beispiel für attentional tunneling das Unglück im Kernkraftwerk „Three Mile Island“ im Jahr 1979 an, bei dem sich die Operateure bei der Problemanalyse zu stark auf einen einzelnen fehlerhaften Indikator fokussierten und andere wichtige Informationen, die eine andere Problemursache nahe legten, außer Acht ließen. Auch bei dem Unglück von Flug AF447, auf das in Kap. 1.1. näher eingegangen wurde, deuten die Analysen darauf hin, dass die Piloten ein nicht zutreffendes mentales Modell von der Situation hatten und unerwartete Signale wie die „Stall“-Warnung nicht wahrnahmen oder nicht genügend in die Problembewertung einbezogen.

Ein adaptives Mensch-Maschine-System kann hierbei Unterstützung bieten. Zum Beispiel ist denkbar, dass es die Salienz und Zugänglichkeit der dargebotenen Informationen problemadäquat anpasst oder per „cueing“ auf nicht beachtete Informationen oder Aufgaben hinweist. Eine Umsetzung dieser Unterstützungsstrategie ist in Fuchs & Schwarz (2017) beschrieben. Für die Nutzerzustandsdiagnose bedeutet dies, dass sie in der Lage sein sollte, Einschränkungen der Aufmerksamkeit, insbesondere die Nichtbeachtung relevanter Informationen oder Aufgaben, zu erkennen.

Vigilanz

Vigilanz bezieht sich auf die Aufrechterhaltung der Aufmerksamkeit über einen längeren Zeitraum. Nach Schmidtke (1989) kann sie definiert werden als eine „*spezifische leistungsbeeinflussende Variable bei Dauerbeobachtungstätigkeiten […] d.h. die Fähigkeit oder die Bereitschaft, gewisse, in der Umwelt in zufälligen zeitlichen Abstand auftretende Veränderungen zu erkennen und auf sie adäquat zu reagieren*“ (HdE Band 2, A-8.3.2, S.1). Studien haben gezeigt, dass die Vigilanz bei Beobachtungstätigkeiten unter reizarmen Bedingungen bereits nach 30 Minuten signifikant abnimmt (Mackworth, 1948; Schmidtke & Micko, 1964). Der Desaktivierungstheorie zufolge kann dieser Vigilanzabfall durch eine adaptationsbedingte Minderung des Neuigkeitswertes einer Situation erklärt werden. Unter Adaptation wird hierbei die Gewöhnung an komplexe Situationen verstanden, die durch ständige Wiederkehr für das Individuum an Neuigkeitswert einbüßen (Schmidtke, 1989). Neurophysiologische Untersuchungen von Moruzzi & Magoun (1949) weisen darauf hin, dass wahrgenommene Informationen nicht nur in der Großhirnrinde verarbeitet werden, sondern auch das Aktivierungszentrum in der Retikulärformation anregen. Dieses regt seinerseits die Großhirnrinde an und schafft die Voraussetzung für kortikale Aktivität. Bleiben diese Anregungseffekte aus oder sinkt infolge Adaption der Weckeffekt, führt dies zu Schlaftendenzen. Auf Basis dieser Erkenntnisse wird empfohlen, die Dauer der ununterbrochenen Beobachtungszeit auf 30 bis maximal 60 Minuten zu begrenzen. Nach dieser Zeit sollte eine völlig andere Tätigkeit ausgeübt werden. Weitere mögliche Maßnahmen, einem Vigilanzrückgang entgegenzuwirken, bestehen nach Schmidtke (1989) in Rückmeldungen über Erfolg oder Misserfolg, Darbietung aufgabenfremder Reize, wie intervallweiser Musik, oder zwischenmenschlichen Kontakten.

Implikationen für die vorliegende Arbeit

Tabelle 13 fasst die Ergebnisse zusammen, die aus den betrachteten Modellen und Theorien für die Berücksichtigung von Aufmerksamkeit und Vigilanz in adaptiven Systemen abgeleitet werden können.

Tabelle 13. Erkenntnisse aus den Theorien und Modellen zur Aufmerksamkeit und Vigilanz

Betrachtete Aspekte	Ergebnisse der Analyse
Relevante Einflussfaktoren	• Ereignisse • Ziele und Erwartungen • Art der Informationsdarbietung • Beanspruchung, negative Emotionen • Monotonie • Situationsbewusstsein (Aufmerksamkeitslenkung durch Erwartung von Informationen, die das mentale Modell bestätigen)
Kritische Ausprägungen	• Nichtbeachtung wichtiger Informationen („attentional tunneling“) • verringerte (Dauer-)Aufmerksamkeit
Wichtige Unterscheidungen	• selektive Aufmerksamkeit • überdauernde Aufmerksamkeit (Vigilanz)
Primäre Auswirkungen	• Situationsbewusstsein
Mögliche Unterstützung durch das adaptive System	• Lenkung der Aufmerksamkeit auf wichtige nicht beachtete Informationen, beschrieben in Fuchs & Schwarz (2017). • Anregung bei Beobachtungstätigkeiten schaffen (vgl. passive aufgabenbezogene Müdigkeit Abschnitt 3.1.4)

3.1.6 Situationsbewusstsein

Das Konzept des Situationsbewusstseins (engl. situation awareness, SA) ist im Bereich der Mensch-Maschine-Interaktion insbesondere in Bezug auf hochautomatisierte Systeme von besonderer Bedeutung (vgl. Abschnitt 1.2). Maßgeblich wurde das Konzept durch die Arbeiten von Mica Endsley geprägt, die in ihre Theorie verschiedene Erkenntnisse integrierte und so ein generisches Modell zum Situationsbewusstsein entwickelte, das als Rahmenwerk für die weitere Forschung fungiert.

Definition

Nach der Definition von Endsley (1988) handelt es sich bei dem Situationsbewusstsein um die Wahrnehmung der Elemente in der Umgebung innerhalb einer bestimmten Zeit und eines bestimmten Raums, das Verstehen ihrer Bedeutung und die Projektion auf ihren Zustand in der nahen Zukunft. Auf Basis dieser Definition werden drei Stufen des Situationsbewusstseins unterschieden, die jeweils aufeinander aufbauen: (1) Die *Wahrnehmung* der gegenwärtigen Situation, (2) das *Verstehen* der gegenwärtigen Situation und (3) die *Projektion* auf den zukünftigen Zustand.

Eng mit dem Situationsbewusstsein verknüpft ist das Konzept des *mentalen Modells*, das als Mechanismus für den Aufbau von Situationsbewusstsein betrachtet werden kann (Rouse & Morris (1986). So definiert Wickens (1996b) basierend auf einer Definition von Dominguez (1994) Situationsbewusstsein als kontinuierliche Aufnahme von Information über ein System oder die Umwelt, die Integration dieser Information mit Vorwissen, um ein stimmiges mentales Bild zu formen und die Verwendung dieses mentalen Bildes, um die weitere Wahrnehmung sowie Antizipation von und Reaktion auf künftige Ereignisse zu steuern.

Einflussfaktoren und Wirkzusammenhänge

Die wesentlichen Wirkzusammenhänge wurden von Endsley (1995) in einem Modell dargestellt, das in Abbildung 12 wiedergegeben ist. Es enthält neben den drei Stufen des Situationsbewusstseins auch die Faktoren, die auf das Situationsbewusstsein Einfluss nehmen. Individuelle Faktoren, wie Fähigkeiten, Erfahrung und Training können dazu führen, dass Individuen unterschiedlich gut in der Lage sind, Situationsbewusstsein aufzubauen. Zudem fungieren die jeweiligen Ziele und Vorannahmen des Operateurs als Filter und beeinflussen die Interpretation der Umgebung. Im Rahmen der Mensch-Maschine-Interaktion nehmen auch Eigenschaften des technischen Systems, wie zum Beispiel die Gestaltung der Benutzungsoberfläche oder der Automationsgrad Einfluss auf das Situationsbewusstsein, da hiervon abhängig ist, ob und in welcher Weise benötigte Informationen bereitgestellt werden. Des Weiteren können Beanspruchung und Stress zu Beeinträchtigungen des Situationsbewusstseins führen. Aufmerksamkeit wird als Voraussetzung für die Wahrnehmung und Verarbeitung von Informationen und damit für den Aufbau von Situationsbewusstsein gesehen, wobei sie auch für die nachgeschalteten Prozesse der Entscheidungsfindung und Handlungsausführung benötigt wird.

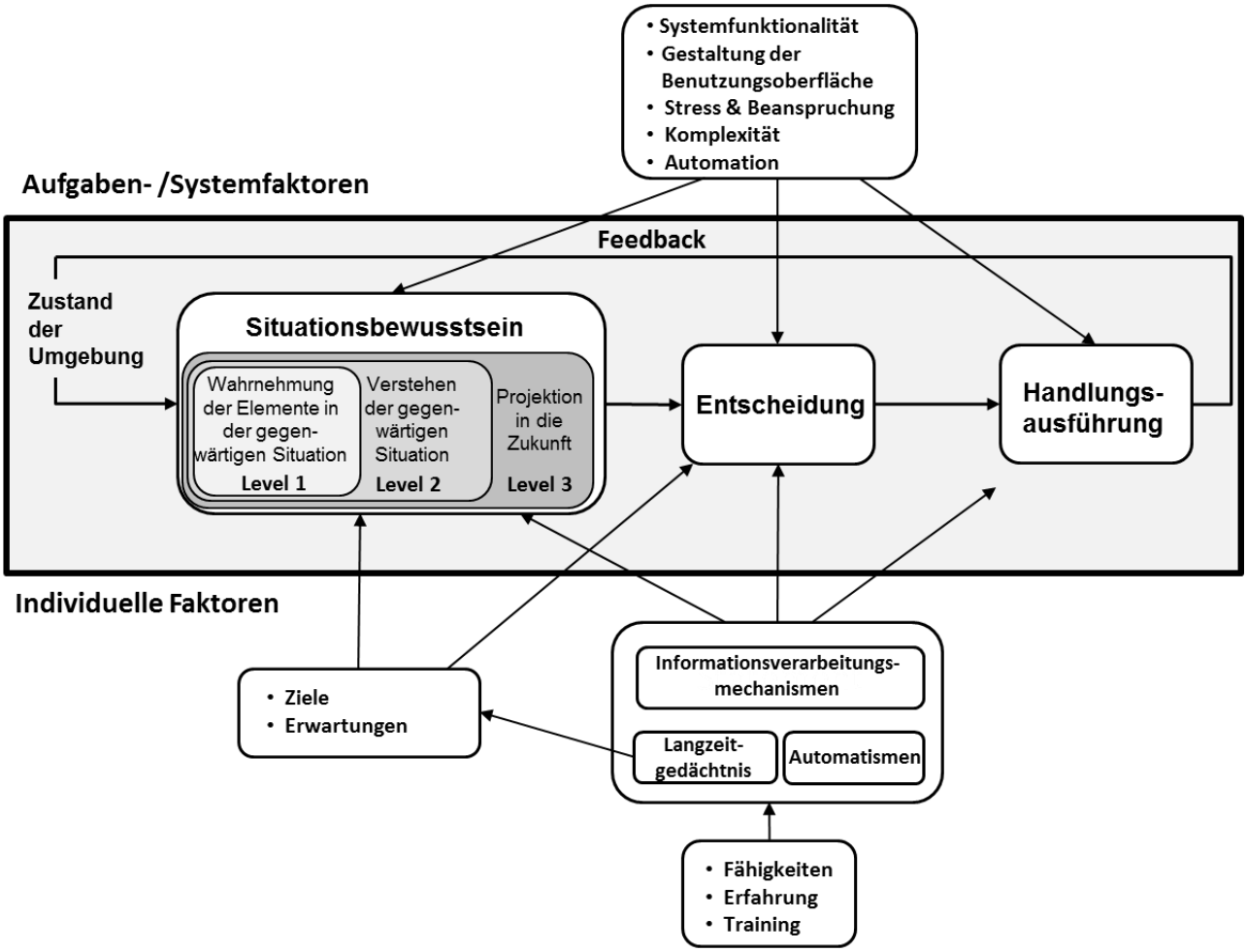


Abbildung 12. Modell zum Situationsbewusstsein nach Endsley (1995)

Auswirkungen des Situationsbewusstseins

Das Situationsbewusstsein bestimmt, welche Entscheidungen getroffen werden. Daraus gehen bestimmte Handlungen hervor, die den Umweltzustand verändern. Ein Verlust des Situationsbewusstseins kann somit zu Fehlentscheidungen und –handlungen führen, die mit gravierenden Konsequenzen verbunden sein können (vgl. Endsley, 1995). Beispiele sind Flugzeugabstürze (Orasanu & Martin, 1998, Endsley, 1995, Hartel, Smith, & Prince, 1991), Friendly Fire (Salmon, Stanton, Walker, & Green, 2006) und Kernkraftwerksunglücke (Flin, O'Connor, & Crichton, 2008), die auf inadäquates Situationsbewusstsein zurückgeführt wurden.

Implikationen für die vorliegende Arbeit

Tabelle 14 fasst die Ergebnisse zusammen, die aus der Analyse von Endsleys Arbeiten für die Berücksichtigung des Situationsbewusstseins in adaptiven Systemen abgeleitet werden können. Die aufgeführten Unterstützungsmöglichkeiten beziehen sich auf einige der von Endsley (1995) selbst genannten Empfehlungen zur Gestaltung von Benutzungsschnittstellen.

Tabelle 14. Erkenntnisse aus den Theorien und Modellen zum Situationsbewusstsein

Betrachtete Aspekte	Ergebnisse der Analyse
Relevante Einflussfaktoren	• Eigenschaften des technischen Systems/Automation • Stress und Beanspruchung • Zustand der Umgebung • Ziele • Fähigkeiten und Erfahrung
Kritische Ausprägungen	• Fehlerhaftes Situationsbewusstsein
Wichtige Unterscheidungen	• Level 1, 2 und 3 des Situationsbewusstseins
Primäre Auswirkungen	• Entscheidungen
Mögliche Unterstützung durch das adaptive System	• Bereitstellen von Informationen, die für den Aufbau von korrektem SA und für die Zielerreichung benötigt werden • Salienz von relevanten Informationen erhöhen und von irrelevanten Informationen verringern • Bei der Vorhersage zukünftiger Ereignisse und Zustände unterstützen

3.2 Wechselwirkungen zwischen den Nutzerzustandsdimensionen

Die Analyse der Theorien und Modelle zu den Nutzerzuständen ergibt, dass jede der betrachteten Dimensionen des Nutzerstands selbst ein multidimensionales Konstrukt darstellt. Dabei zeigt sich, dass zwischen diesen Zuständen nicht nur vielfältige Wechselwirkungen bestehen, sondern dass bestimmte Ausprägungsformen teilweise auch ineinander übergehen und nicht klar voneinander abzugrenzen sind. Zum Beispiel kann auf Basis der theoretischen Analyse angenommen werden, dass geringe Beanspruchung, passive aufgabenbezogene Müdigkeit und Langeweile ähnliche mentale Zustände repräsentieren, die wiederum eng mit der Vigilanz verbunden sind. Hohe Beanspruchung steht in engem Zusammenhang mit der aktiven aufgabenbezogenen Müdigkeit und weist Überschneidungen mit Stress, als einer Ausprägungsform des emotionalen Zustands, auf. Der Begriff der „emotionalen Beanspruchung" verdeutlicht diese konzeptionelle Ähnlichkeit.

Es zeigte sich auch, dass Aufmerksamkeit und Situationsbewusstsein eng miteinander verwoben sind. So bezieht sich das Level-1-SA auf die Wahrnehmung der Elemente in der Umgebung. Dies ist nur dann möglich, wenn die Aufmerksamkeit auf diese Informationen gerichtet wird.

Motivation bezieht sich auf die Anstrengungsbereitschaft und wirkt sich somit auf die Beanspruchung aus. Motivationale Zustände stehen auch im Zusammenhang mit emotionalen Zuständen. So können emotionale Zustände die Handlungs- und Anstrengungsbereitschaft erhöhen (z.B. positive Emotionen, Eustress) oder auch vermindern (z.B. Langeweile, Frustration). Die Analysen weisen außerdem darauf hin, dass sich Beanspruchung, Müdigkeit, Motivation und emotionale Zustände auf die Aufmerksamkeit und das Situationsbewusstsein auswirken.

Zu beachten ist, dass die theoretische Analyse zwar kausale Abhängigkeiten zwischen den verschiedenen Dimensionen nahe legt, diese aber selten empirisch untersucht wurden. Edwards (2013) analysierte die Zusammenhänge zwischen verschiedenen mentalen Zuständen (Beanspruchung, Stress, Müdigkeit, Aufmerksamkeit und Situationsbewusstsein) und der Leistung bei Fluglotsen. Sie konstatiert, dass empirische Untersuchungen hierzu meist korrelativer Natur sind und somit keine Aussagen über Kausalzusammenhänge zulassen. Sie konnte jedoch belegen, dass zwischen den Zuständen Wechselwirkungen bestehen und je nach Kombination die Leistung unterschiedlich stark beeinflusst wird. Insgesamt bestätigen diese Ergebnisse ebenfalls die Notwendigkeit den Nutzerzustand nicht eindimensional, sondern multidimensional als Zusammenwirken von verschiedenen mentalen Zuständen zu betrachten. Da die Zusammenhänge zwischen den Dimensionen des Nutzerzustands, wie beschrieben vielfältig aber noch nicht vollständig untersucht sind, wird in der vorliegenden Arbeit davon ausgegangen, dass zwischen allen sechs betrachteten Nutzerzuständen Wechselwirkungen bestehen können. Eine entsprechende Darstellung des multidimensionalen Nutzerzustands ist in Abbildung 13 wiedergegeben.

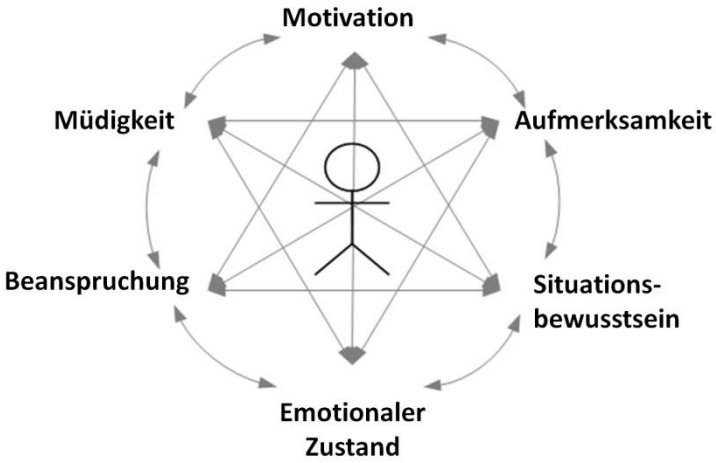


Abbildung 13. Darstellung des multidimensionalen Nutzerzustands

3.3 Einflussfaktoren auf Nutzerzustand und Informationsverarbeitung

In den Modellen und Theorien zu den sechs Dimensionen des Nutzerzustands konnten verschiedene Einflussfaktoren identifiziert werden, die sich auf den Nutzerzustand und die menschliche Informationsverarbeitung auswirken, und somit im Rahmen einer ganzheitlichen Bewertung des Nutzerzustands berücksichtigt werden sollten. Dieser Abschnitt stellt die verschiedenen

Einflussfaktoren in Kategorien geordnet näher vor. Unterschieden wird dabei zwischen individuellen Faktoren, Aufgaben, Umgebung, Kontext, Eigenschaften des technischen Systems sowie Zielen und Ereignissen.

3.3.1 Individuelle Faktoren

Wie sich in den Theorien und Modellen zum Nutzerzustand zeigte, können verschiedene individuelle Faktoren den Nutzerzustand beeinflussen, z.B. sind Kenntnisse, Fähigkeiten und Erfahrungen wesentliche Determinanten der Beanspruchung; Schlaf wirkt sich auf die Müdigkeit aus und persönliche Ziele beeinflussen die Motivation. Hinsichtlich der mentalen Beanspruchung wird in der ISO-Norm DIN EN 10 075-1 (2000) zwischen relativ überdauernden stabilen Faktoren und variablen Faktoren unterschieden, die sich von Tag zu Tag oder auch im Laufe eines Tages verändern können (vgl. Abschnitt 3.1.1). Veltman, Hockey, Schlegel, Fraser, & Burov (2004) differenzieren diesbezüglich zwischen den Konzepten *Background State* und *Baseline State*, die sie in den Zusammenhang mit dem Operator Functional State (OFS, siehe auch Abschnitt 2.2.7) stellen. Für die vorliegende Arbeit wird diese Unterscheidung übernommen, um die Wirkzusammenhänge zwischen individuellen Faktoren und Nutzerzustand zu verdeutlichen. Nach Veltman et al. (2004) bezeichnet der Background State den durchschnittlichen unbeanspruchten Zustand des Operateurs, der losgelöst ist von jeglichen Anforderungen und Zielen. Der Background State ergibt sich aus den Eigenschaften des Operateurs und wird als zeitlich stabil angesehen. Als wichtige Determinanten können hierbei Persönlichkeitseigenschaften, Fähigkeiten, Kenntnisse, der Erfahrungsgrad, die Konstitution und Bewältigungsstrategien angesehen werden.

Als Baseline State bezeichnen Veltman et al. (2004) demgegenüber den aktuellen nicht belasteten Zustand des Operateurs, der vorliegt, bevor der Operateur mit seiner operativen Tätigkeit beginnt. Dieser wird durch variable individuelle Faktoren (z.B. Stimmung, Wohlbefinden) oder vorangegangene Tätigkeiten (Schlaf, Arbeitstätigkeit) beeinflusst. Der Baseline State setzt sich somit sowohl aus den stabilen als auch den temporär variablen Eigenschaften des Operateurs zusammen. Der gegenwärtige Nutzerzustand ist schließlich das Resultat des Baseline State und den auf den Nutzer einwirkenden Merkmalen der Anforderungssituation. Die Zusammenhänge sind in Abbildung 14 skizziert.

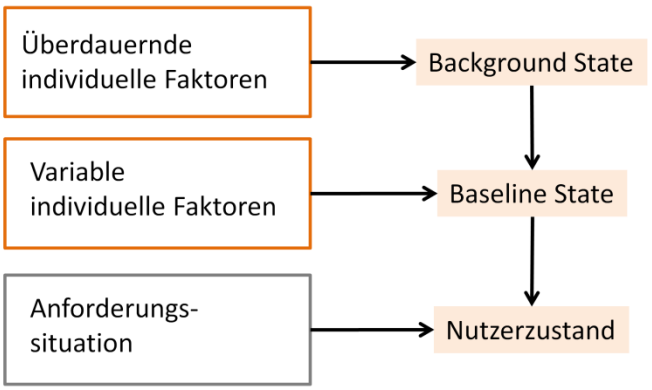


Abbildung 14. Wirkzusammenhänge zwischen individuellen Faktoren und dem Nutzerzustand basierend auf Veltman et al. (2004)

3.3.2 Aufgaben

Aufgaben lassen sich auf vielfältige Weise beschreiben und klassifizieren. In den Modellen zu den Dimensionen des Nutzerzustands hat sich herausgestellt, dass insbesondere die Aufgabenart, die Dauer und die Schwierigkeit wesentlichen Einfluss auf den Nutzerzustand ausüben.

Hinsichtlich der Art der Aufgabe ist es sinnvoll, entsprechend dem Modell der multiplen Ressourcen von Wickens (1984), zwischen Wahrnehmungsaufgaben, kognitiven Aufgaben und (psycho-)motorischen Aufgaben zu unterscheiden. Bei Wahrnehmungsaufgaben sollte zudem differenziert werden, inwiefern die Aufgabe die jeweiligen Sinneskanäle (z.B. den visuellen oder den auditiven Kanal) auslastet. Zudem sollte berücksichtigt werden, ob Aufgaben parallel bearbeitet werden müssen und wenn ja, wie sie kodiert sind, und ob die Ausführung vokal oder manuell erfolgt. Die Aufgabendauer spielt insbesondere bei Überwachungsaufgaben in Hinblick auf die Vigilanz eine wesentliche Rolle. Hinsichtlich der Beanspruchung sind die Aufgabenschwierigkeit, die zur Verfügung stehende Zeit und die Menge der Aufgaben von Belang. Aufgabendauer und -schwierigkeit beeinflussen außerdem die aufgabenbezogene Müdigkeit.

3.3.3 Umgebung

Umgebungsfaktoren beziehen sich auf Zustände in der Umwelt. Zu den Umgebungsfaktoren, die als Stressor wirken und maßgeblich die Leistungsfähigkeit beeinflussen können, zählt die *Umgebungstemperatur*. Verschiedene Studien stellten fest, dass sowohl besonders hohe als auch niedrige Temperaturen bei verschiedenen kognitiven Aufgaben zu Leistungsminderungen führen. Einer metaanalytischen Untersuchung von Pilcher, Nadler, & Busch (2002) zufolge, in der 515 Effektgrößen aus 22 Studien berücksichtigt wurden, treten die größten Leistungseinbußen bei Temperaturen über 90°F (~32°C) und unter 50°F (~10°C) auf.

Als ein weiterer Stressor ist *Lärm* anzusehen. Hockey (1986) führt aus, dass Lärm die Effektivität des Kurzzeitgedächtnisses stört. Die Ausführung von Aufgaben, die das Abspeichern und Abrufen von Inhalten aus dem Kurzzeitgedächtnis beinhalten (z.B. Problemlösungsaufgaben, arithmetische Aufgaben), wird somit beeinträchtigt. Trimmel & Poelzl (2006) stellten fest, dass Hintergrundgeräusche selbst niedriger Intensität die kortikalen Ressourcen reduzieren und die Leistung in Bezug auf räumliche Aufmerksamkeit verschlechtern. In einigen Studien, die in Hockey (1970) aufgeführt sind, wurde jedoch auch eine Verbesserung der Leistung bei Vigilanzaufgaben unter Lärmbedingungen festgestellt. Hockey (1970) erklärt die widersprüchlichen Ergebnisse damit, dass der Effekt, den Lärm verursacht, von der Komplexität der Aufgabe abhängt. So erhöhe Lärm das Arousal und führe zu einer höheren Aufmerksamkeitsselektivität (Fokussierung auf bestimmte Stimuli unter Vernachlässigung von anderen, vgl. Abschnitt 3.1.5). Diese wirke sich bei einfachen Aufgaben positiv aus. Bei komplexen Aufgaben, in denen eine Vielzahl an Informationen wahrgenommen werden muss, würde diese Wahrnehmungseinengung die Leistung hingegen verschlechtern.

Des Weiteren sind auch *Lichtverhältnisse* zu berücksichtigen. So kann helles Umgebungslicht die Müdigkeit reduzieren, indem es die Melatoninproduktion hemmt (Gawron et al., 2001). Stefani & Krüger (2013) konnten nachweisen, dass selbst die Bildschirmhelligkeit die innere Uhr beeinflusst und sich auf Müdigkeit und Gedächtnisleistung auswirkt.

Je nach Anwendungskontext müssen auch spezifische Umgebungsfaktoren betrachtet werden. So wird die Leistungsfähigkeit auf Schiffen durch *Seegang* beeinflusst. Schoeffel (1987) berichtet, dass es bei Sonarbeobachtern bei hohen Seegangsstärken während einer 14-tägigen Fahrt zu deutlichen Einbrüchen der Vigilanzleistungen kam. Bei Düsenjetpiloten muss die Einwirkung von Trägheitskräften („G-Kraft“) berücksichtigt werden.

3.3.4 Kontext

Unter dem Kontext können unterschiedliche Bedingungen subsumiert werden, die Einfluss darauf haben, wie der Operateur auf situative und aufgabenbezogene Anforderungen reagiert. Im Folgenden werden einige Beispiele für relevante Kontextfaktoren aufgeführt.

Ein wichtiger Aspekt, der insbesondere für die Motivation relevant ist, ist die Unterscheidung, ob es sich um eine *Trainings- oder eine Einsatzsituation* handelt. Veltman & Jansen (2004) bemerken hierzu, dass die Anstrengungsbereitschaft beim Fliegen eines realen Flugzeugs deutlich höher ist als im Simulator. Wenn es bei dem Flug im Simulator jedoch darum geht, als Pilot ausgewählt zu werden, ist die mentale Anstrengung genauso hoch wie in einem realen Flugzeug (Veltman, 2002).

Zudem zeigten die Hawthorne-Studien (Roethlisberger & Dickson, 1939), dass die Leistung in starkem Maße vom *sozialen Kontext* und von der Anerkennung z.B. durch Kollegen und Vorgesetzte abhängt. Der so genannte *Hawthorne-Effekt* weist außerdem darauf hin, dass experimentelle Untersuchungen ebenfalls einen besonderen Kontext darstellen, der sich auf die Leistungsbereitschaft der Versuchsteilnehmer auswirken kann. Diese kontextbezogenen Effekte können in Hinblick auf Maslows Theorie (Maslow, 1943) durch unterschiedlich stark ausgeprägte Bedürfnisse in den jeweiligen Kontextsituationen erklärt werden oder nach Porter & Lawler (1968) durch unterschiedliche Erwartungswerte (vgl. Abschnitt 3.1.3).

Ein weiterer Kontextfaktor, der sich insbesondere auf die Müdigkeit auswirkt, ist die *Tageszeit*. Bedingt durch den circadianen Rhythmus ist der menschliche Organismus in Abhängigkeit von der Tageszeit unterschiedlich anfällig für schlafbezogene Müdigkeit (vgl. Abschnitt 3.1.4). In diesem Zusammenhang stellte sich heraus, dass die Leistung zwischen 2 und 6 Uhr nachts schlechter ist als zu anderen Tageszeiten (Johnson & Naitoh, 1974; Meddis, 1982; Akerstedt, 1990).

3.3.5 Eigenschaften des technischen Systems

Wie in Kapitel 1.2 bereits ausgeführt wurde, ist insbesondere der *Automationsgrad* des technischen Systems ein wichtiger Faktor, der sich auf den Nutzerzustand (z.B. Beanspruchung, Aufmerksamkeit, Situationsbewusstsein) und damit auf die Leistungsfähigkeit des Operateurs auswirken kann. Als weitere wichtige Eigenschaften des technischen Systems sind die Zuverlässigkeit, die *Komplexität* und die *Benutzungsfreundlichkeit* („Usability“) zu nennen (vgl. Kap. 1.2), die auch von Endsley (1995) als Einflussfaktoren auf das Situationsbewusstsein aufgeführt werden.

Eigenschaften des technischen Systems können nach Feigh et al. (2012) auch als Trigger für die Adaptierung herangezogen werden. Adaptierungsstrategien werden dabei in Abhängigkeit von Zustand oder Modus des technischen Systems ausgelöst. Im Automobil bezieht sich der Zustand

zum Beispiel auf Position, Geschwindigkeit und Beschleunigung. Der Modus bezieht sich auf bestimmte Systemverhaltensweisen, die der Operateur überwachen oder auswählen kann.

3.3.6 Ziele

Ziele bestimmen, wie in Abschnitt 3.1.3 näher ausgeführt wurde, wesentlich die Motivation und das Verhalten des Menschen. Sie können in einem Top-Down-Prozess die Aufmerksamkeit lenken (vgl. Abschnitt 3.1.5) und sind auch eine wichtige Grundlage für den Aufbau von Situationbewusstsein (vgl. Endsley 1995, Abschnitt 3.1.6). In Mensch-Maschine-Systemen kann der Mensch verschiedene Ziele verfolgen. Ein wichtiges Ziel besteht zumeist darin, Sicherheit zu gewährleisten. Im Transportbereich könnten weitere Ziele darin bestehen, schnell an den Bestimmungsort zu gelangen oder unter ökonomischen Gesichtspunkten wenig Treibstoff zu verbrauchen. Wenn Ziele miteinander kompatibel sind, können sie zur gleichen Zeit verfolgt werden (z.B. Sicherheit gewährleisten und wenig Treibstoff verbrauchen; wenn sie nicht kompatibel sind (z.B. schnell ankommen und wenig Treibstoff verbrauchen), bestimmt die Priorität der Ziele, welches Ziel (zuerst) verfolgt wird (vgl. Endsley, 1995).

Ziele können außerdem hierarchisch in über- und untergeordnete Ziele gegliedert werden (Austin & Vancouver, 1996). So können übergeordnete Ziele (z.B. den Bestimmungort erreichen) in konkrete Handlungsziele, die auch als Aufgaben bezeichnet werden können, untergliedert werden (z.B. Bestimmungsort im Navigationssystem eingeben). Im Kontext adaptiver Systemgestaltung sollte das technische System Kenntnis über die jeweiligen Ziele und damit verbundenen Aufgaben des Mensch-Maschine-Systems sowie ihre Priorität erhalten, um den Operateur bei der Zielerreichung adäquat unterstützen zu können (vgl. Fuchs et al., 2007).

3.3.7 Ereignisse

Ereignisse können den Nutzerzustand in Abhängigkeit dessen, wie sie durch den Operateur bewertet werden, unterschiedlich beeinflussen. Wie sich in Abschnitt 3.1.2 zeigte, ist die Bewertung eines Ereignisses bezüglich der Bedeutsamkeit, dem Stressniveau und der Kontrollierbarkeit ausschlaggebend dafür, wie sich das Ereignis auf den emotionalen Zustand und die Leistung auswirkt. Potenziell kritisch sind dabei insbesondere unerwartete Störfälle („emergency situations“, Wickens, 1996a). Im Fallbeispiel in Kapitel 1 wurde deutlich, dass der Verlust der Geschwindigkeitsinformation in Kombination mit dem Ausfall des Autopiloten mit hoher Bedeutsamkeit, hohem Stressniveau und geringer Kontrollierbarkeit verbunden war. Dies löste bei den Piloten eine Reihe kritischer Nutzerzustände aus und führte letztendlich zu einem völligen Verlust der Kontrolle über das Flugzeug. Ein adaptives System könnte bei Auftreten von unerwarteten Störfällen insbesondere in Hinblick auf die Kontrollierbarkeit unterstützen, indem es Handlungsmöglichkeiten aufzeigt oder dem Operateur durch Übernahme von Aufgaben ermöglicht, seine kognitiven Ressourcen der Ursachenanalyse und Behebung des Störfalls zu widmen.

3.4 Ableitung eines generischen Modells zum Nutzerzustand

Auf Basis der zuvor beschriebenen Modelle zu den betrachteten Dimensionen des Nutzerzustands und den sich daraus ergebenden Erkenntnissen wurde ein Modell entwickelt, das in generischer Form beschreibt, welche Faktoren Einfluss auf die Entstehung und Regulierung des aktuellen Nutzerzustands nehmen, und welche Reaktionen aus dem Nutzerzustand resultieren können (Abbildung 15). Das Modell kann somit als Grundlage zur Identifizierung der für eine Nutzerzustandsdiagnose relevanten Faktoren bei verschiedenen Anwendungsdomänen verwendet werden. Im Folgenden werden die im Modell dargestellten Komponenten und Wirkzusammenhänge näher erläutert. Zur besseren Nachvollziehbarkeit sind diese im Modell mit Buchstaben markiert, auf die im Text in Klammern verwiesen wird.

3.4.1 Komponenten des Modells

Analog zu dem in Kapitel 1 vorgestellten MMI-Modell (s. Abbildung 1) sind Faktoren und Prozesse innerhalb des Operateurs *orange* und die auf den Operateur einwirkenden externen Faktoren *grau* dargestellt. Letztere beschreiben die jeweilige Anforderungssituation (A). Diese ist gegliedert in Aufgabenmerkmale, Umgebungs- und Kontextfaktoren, Eigenschaften des technischen Systems, Zielvorgaben und Ereignisse (vgl. Abschnitte 3.3.2 - 3.3.7). Die Anforderungssituation beeinflusst den Nutzerzustand direkt (a), z.B. beeinflussen Tageszeit und Aufgabendauer die Müdigkeit.

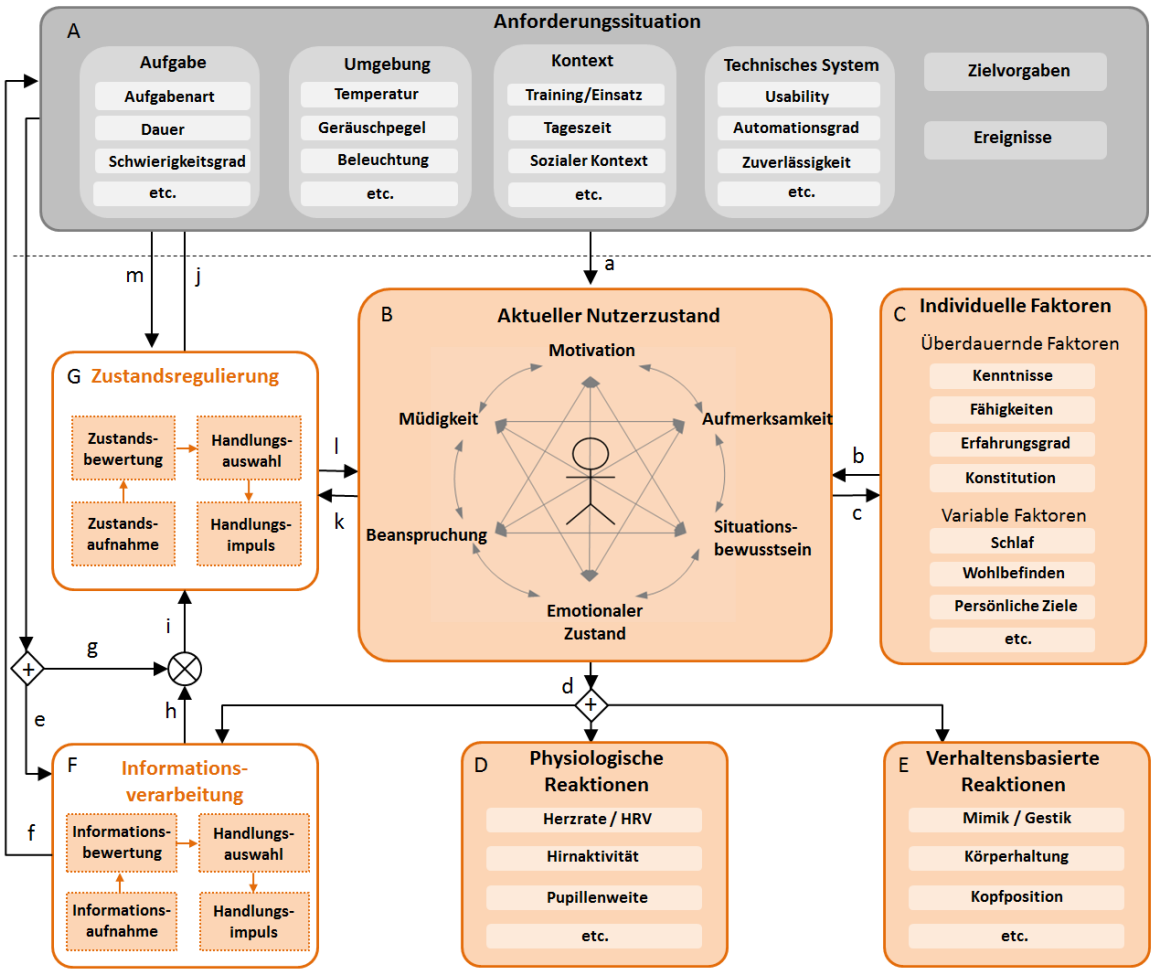


Abbildung 15. Generisches Modell zu Einflussfaktoren und Auswirkungen des Nutzerzustands

In der Mitte des Modells ist der multidimensionale Nutzerzustand abgebildet, der aus den sechs Zustandsdimensionen und ihren Wechselwirkungen resultiert (B). Rechts daneben befinden sich die individuellen Faktoren (C), die den Nutzerzustand direkt beeinflussen können (b). Sie sind entsprechend den Ausführungen in Abschnitt 3.3.1 unterteilt in überdauernde Faktoren und variable Faktoren, die den Background- bzw. Baselinestate darstellen (vgl. Abbildung 14). Der Nutzerzustand kann allerdings auch zu Veränderungen bei den individuellen Faktoren führen (c). So kann Müdigkeit in Folge von Schlafentzug langfristig den Gesundheitszustand beeinträchtigen (Miller et al., 2011). Der Nutzerzustand wirkt sich zudem auf verschiedene weitere Komponenten aus (d), die unterhalb des Modells dargestellt sind. Dazu zählen Reaktionen in Physiologie (D) und Verhalten (E), die zur Diagnose des Nutzerzustands herangezogen werden können (vgl. Abschnitt 2.3.3). Außerdem nimmt der Nutzerzustand, wie in den vorangegangenen Abschnitten 3.1.1 - 3.1.6 gezeigt wurde, Einfluss auf die menschliche Informationsverarbeitung (F). Diese wirkt sich wiederum, wie in Abschnitt 3.4.3 beschrieben wird, auf die Prozesse der Zustandsregulierung (G) aus.

3.4.2 Der Informationsverarbeitungsprozess

Für die Darstellung des Informationsverarbeitungsprozesses wurden die vier Stufen der Informationsverarbeitung nach Parasuraman, Sheridan & Wickens (2000) zugrunde gelegt:

1. *Informationsaufnahme:* Über die Sinnesorgane werden Informationen aus verschiedenen Quellen der Umwelt aufgenommen (e). Diese Stufe beinhaltet die Vorverarbeitung sensorischer Informationen vor der bewussten Wahrnehmung und die selektive Aufmerksamkeitszuwendung zu bedeutsamen Informationen.
2. *Bewertung:* Auf dieser Stufe werden die Informationen bewusst wahrgenommen und im Arbeitsgedächtnis zusammen mit Informationen aus dem Langzeitgedächtnis verarbeitet und bewertet.
3. *Handlungsauswahl:* Auf Basis der kognitiven Verarbeitung werden Handlungsoptionen generiert und ausgewählt.
4. *Handlungsimpuls:* Entsprechend der in Stufe 3 getroffenen Entscheidung wird eine Handlung initiiert. In Bezug auf die Aufgabenbewältigung können diese Aktionen als Leistung bezeichnet werden (f). Das Resultat der Handlungen wird wiederum wahrgenommen und bewertet (e,g).

3.4.3 Der Zustandsregulierungsprozess

Eine weitere wichtige Komponente ist die Zustandsregulierung (G), deren Wirkzusammenhänge im Modell basierend auf den Ausführungen zur Anstrengungsregulierung von Veltman & Jansen (2006) dargestellt wurden (vgl. Abschnitt 3.1.3). Demnach finden Ist-Soll-Vergleiche zwischen benötigter Leistung und aktueller Leistung (h) statt, wobei die Diskrepanz zwischen Ist und Soll die Zustandsregulierung steuert. So kann eine Diskrepanz dazu führen, dass mehr Anstrengung aufgebracht wird, um die Leistung zu steigern (i). Die Ist-Soll-Diskrepanz kann jedoch auch

dadurch reduziert werden, dass die Aufgaben und Ziele angepasst werden, z.B. indem weniger relevante Aufgaben vernachlässigt werden oder eine weniger gute Leistung akzeptiert wird (j).

Das Modell von Veltman & Jansen (2006) berücksichtigt jedoch nicht, dass Zustandsregulierungsprozesse auch in Hinblick auf die übrigen Dimensionen des Nutzerzustands stattfinden können. Beispielsweise kann der Mensch Müdigkeit entgegenwirken, indem er eine Ruhepause einlegt oder Stimulantien, wie Koffein, zu sich nimmt (vgl. Abschnitt 3.1.4). In der vorliegenden Arbeit wird die menschliche Selbstregulierung daher umfassender als ein Prozess betrachtet, der das Ziel verfolgt, aversive Zustände, die das Wohlbefinden und die Leistung beeinträchtigen, zu vermeiden. Analog zum Informationsverarbeitungsprozess wird dieser Prozess im Modell über die Stufen: *Zustandsaufnahme*, *Zustandsbewertung*, *Handlungsauswahl* und *Handlungsausführung* beschrieben. In Bezug auf Müdigkeit würde der Prozess beispielsweise so ablaufen, dass der Organismus auf Stufe 1 zunächst wahrnimmt, dass er müde ist (k). Er bewertet daraufhin auf Stufe 2, ob Aktionen notwendig sind, um den Zustand zu verändern. Als nächstes werden Möglichkeiten abgewogen, mit denen die Müdigkeit reduziert werden kann (z.B. Kaffee trinken, Ruhepause einlegen) und es wird eine Handlungsmöglichkeit ausgewählt (Stufe 3). Die ausgewählte Handlung wird auf Stufe 4 daraufhin ausgeführt (l). Anschließend erfolgt in einer Rückkopplungsschleife eine Bewertung, ob die Handlung erfolgreich war, das heißt im Beispiel, ob die Müdigkeit erfolgreich reduziert werden konnte (k).

Die Art der Zustandsregulierung ist auch von der Anforderungssituation abhängig (m). Zum Beispiel ist das Einlegen einer Ruhepause nur dann möglich, wenn die Situation dies zulässt. Zudem kann der Prozess der Zustandsregulierung, und somit die Fähigkeit zur Anpassung an sich ändernde Anforderungen, durch Stressoren in der Umgebung, wie Temperatur, Lärm, Fliehkräfte beeinträchtigt werden (Veltman & Jansen, 2004).

3.5 Resümee

Das in Abschnitt 3.4 beschriebene generische Modell zum Nutzerzustand enthält die wesentlichen Faktoren, die für die Entstehung und die Auswirkungen des multidimensionalen Nutzerzustands – und damit für eine multifaktorielle Bewertung des Nutzerzustands – von Bedeutung sein können. Diese Erkenntnisse wurden ausschließlich literaturbasiert gewonnen. Um die im Modell enthaltenen Wirkzusammenhänge empirisch zu überprüfen, werden in den folgenden Kapiteln 4 und 5 verschiedene Einflussfaktoren und Auswirkungen des Nutzerzustands für ausgewählte Nutzerzustandsdimensionen experimentell untersucht. Dabei werden auch die aus Kapitel 2 gewonnenen Anforderungen (z.B. Untersuchung der zeitlichen Stabilität, Analyse auf Individualebene) berücksichtigt.

Durch die umfassende Betrachtung der mit dem Nutzerzustand in Zusammenhang stehenden Faktoren, dient das Modell in dieser Dissertation als Grundlage für die Konzeption einer multifaktoriellen Echtzeitdiagnose des Nutzerzustands. Die generische Betrachtung von Faktoren, die als Indikatoren des Nutzerzustands dienen können, bietet zudem einen Orientierungsrahmen für eine anwendungsspezifische Auswahl geeigneter Indikatoren des Nutzerzustands im Rahmen der Umsetzung. Konzeption, exemplarische Umsetzung und Validierung der auf Basis dieser Erkenntnisse entwickelten Diagnosefunktion werden in den Kapiteln 6 und 7 beschrieben.

4 Experiment 1 – Untersuchung einer multifaktoriellen Nutzerzustandsbewertung

Bei den Analysen zum Stand der Forschung (Kapitel 2) und zu den psychologischen Theorien und Modellen zum Nutzerzustand (Kapitel 3) wurden verschiedene Erkenntnisse für eine multifaktorielle Nutzerzustandsbewertung in adaptiver Mensch-Maschine-Interaktion gewonnen. Um die literaturbasierten Erkenntnisse empirisch zu validieren, wurden in der vorliegenden experimentellen Untersuchung verschiedene Dimensionen des Nutzerzustands multifaktoriell auf Basis einer komplexen Überwachungsaufgabe bewertet.

4.1 Forschungsziele

Konkret sollten die folgenden zwei Forschungsziele empirisch untersucht werden:

1) Da die Echtzeitdiagnose in der Lage sein sollte, Ursachen für Nutzerzustandsänderungen zu identifizieren (vgl. Anforderung in Abschnitt 2.4.5), ist zu untersuchen, inwiefern verschiedene Einflussfaktoren auf den Nutzerzustand Veränderungen des Nutzerzustands hervorrufen und erklären können. Im generischen Modell zum Nutzerzustand sind diese als Merkmale der Anforderungssituation und individuelle Faktoren dargestellt (vgl. Abbildung 15, Abschnitt 3.4).
2) Bezugnehmend auf die Analysen zu den Erfassungsmethoden des Nutzerzustands in Abschnitt 2.3 sollte überdies empirisch geprüft werden, ob ausgewählte physiologische und verhaltensbasierte Maße geeignet sind, Veränderungen des Nutzerzustands anzuzeigen.

Wie bei der Analyse von Anforderungen an die Echtzeitdiagnose in Kapitel 2 dargelegt wurde, ist für eine personenspezifische Diagnose die Betrachtung der Ergebnisse auf individueller Ebene notwendig (vgl. Abschnitt 2.4.3). Die Analysen zu den Erfassungsmethoden erfolgten daher nicht nur auf Gruppenebene sondern auch auf Individualebene.

Abbildung 16 zeigt eine vereinfachte Version des in Kapitel 3 beschriebenen generischen Nutzerzustandsmodells (vgl. Abbildung 15, S. 65), in der die im vorliegenden Experiment betrachteten Komponenten hervorgehoben sind. Im Unterschied zum generischen Modell sind in diesem Modell die Leistung und die subjektive Bewertung als messbare Komponenten aufgeführt. Dazu sei angemerkt, dass sich die Leistung aus den im generischen Modell dargestellten Prozessen der Informationsverarbeitung ergibt (vgl. Abschnitt 3.4.2). Die subjektive Bewertung des Nutzerzustands kann als Bestandteil des im generischen Modell aufgeführten Zustandsregulierungsprozesses beim Menschen betrachtet werden (vgl. Abschnitt 3.4.3). Subjektive Maße und Leistungsmaße wurden zwar für die zu entwickelnde Echtzeitdiagnose als weniger geeignet bewertet (vgl. Abschnitt 2.4), sie können jedoch als Vergleichsmaße zur Validierung der Diagnosefähigkeiten der physiologischen und verhaltensbasierten Maße herangezogen werden und dazu beitragen, die Bedeutung der Einflussfaktoren (Merkmale der Anforderungssituation und individuelle Faktoren) zu bewerten.

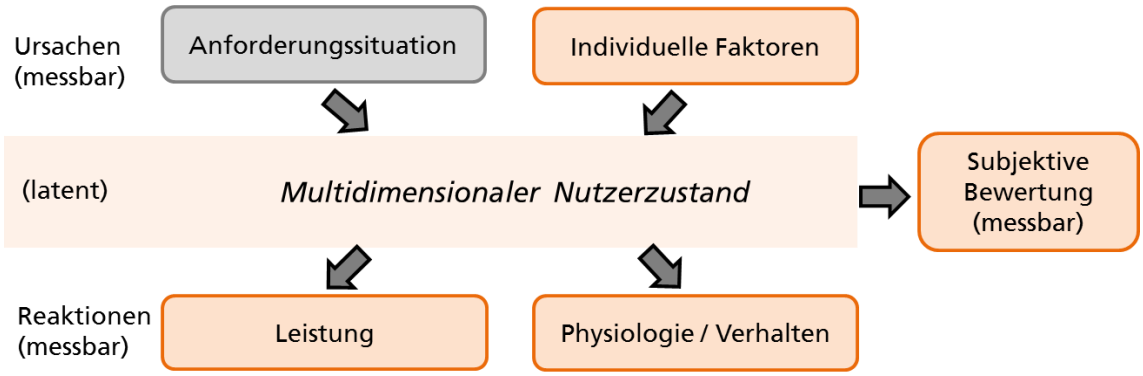


Abbildung 16. Modell mit den in Experiment 1 untersuchten Komponenten zur Nutzerzustandsbewertung

4.2 Methodisches Vorgehen

Um die Forschungsziele zu untersuchen, wurden zur Reduzierung der Komplexität drei der sechs in Kapitel 3 beschriebenen Nutzerzustandsdimensionen ausgewählt und experimentell variiert:

- die *mentale Beanspruchung* (durch Variation der mentalen Belastung),
- die *Frustration* als Ausprägungsform des *emotionalen Zustands* und die
- (selektive) *Aufmerksamkeit*.

Im Folgenden werden die Experimentalaufgabe, das Versuchsdesign, die erfassten unabhängigen und abhängigen Variablen, die konkreten Forschungshypothesen sowie Stichprobe und Ablauf der Untersuchung dargestellt. Außerdem wird auf Besonderheiten bei der Datenaufbereitung und -auswertung eingegangen.

4.2.1 Experimentalaufgabe

Für die experimentelle Untersuchung wurde eine praxisbezogene Aufgabe aus dem Bereich der radargestützten Luftraumüberwachung ausgewählt. Auf standardisierte Laboraufgaben wurde bewusst verzichtet, da diese die operativen Bedingungen in der Realität oft nicht adäquat widerspiegeln (vgl. Abschnitt 2.4.8).

Bei dem verwendeten Experimentalsystem handelt es sich um einen Demonstrator, der nach dem Vorbild eines realen Ausbildungssystems entwickelt wurde. Mit Hilfe einer Simulationssoftware (STAGE von Presagis Inc.) ist eine anwendungsnahe Bearbeitung von Szenarien der Luftraumüberwachung möglich. Damit die Experimentalaufgabe auch von Personen bearbeitet werden kann, die keine Vorkenntnisse in diesem Anwendungsbereich haben, waren für das Experiment einige Modifikationen (im Sinne von Vereinfachungen) der Originalaufgabe notwendig. Dennoch sollte die Experimentalaufgabe die kognitiven Anforderungen der Realaufgabe weitgehend abbilden.

Tabelle 15 stellt die wesentlichen Merkmale der Realaufgabe dar und zeigt, wie diese im Experiment durch eine Primär- und Sekundäraufgabe respräsentiert sind. Zur Leistungsbewertung wurde eigens für das Experiment ein Punktesystem entwickelt. Primär- und Sekundäraufgabe sowie das Punktesystem werden nachfolgend näher erläutert.

Tabelle 15. Merkmale der Realaufgabe und Repräsentation durch die Experimentalaufgabe

Art der mentalen Belastung	Realaufgabe	Experimentalaufgabe
Visuell räumlich	Überwachung der Flugsicherheit: Kontakte müssen innerhalb ihrer Luftraumgrenzen bleiben und Sicherheitsabstände einhalten	Primäraufgabe: Überwachung und Steuerung von Luftkontakten, die sich zufallsgesteuert auf einem (simulierten) Radarbildschirm bewegen
Auditiv verbal	Funkkontakt mit Piloten auf Englisch	Sekundäraufgabe: Lösen von Rechenaufgaben, die auditiv in Englisch präsentiert werden .
Kognitiv	Kursberechnungen zur Durchführung von Abfangmanövern	

Primäraufgabe: Überwachung und Steuerung von Luftkontakten

Die Primäraufgabe besteht darin, zu überwachende Luftkontakte, die sich zufallsgesteuert durch einen simulierten Luftraum bewegen, durch Navigationsanweisungen in vorgegebenen Luftraumgrenzen zu halten. Kontakte dürfen dabei nicht miteinander oder mit Fremdverkehr, der durch die Gebiete fliegt, kollidieren. Die Benutzungsoberfläche mit dem simulierten Radarbild ist in Abbildung 17 wiedergegeben.

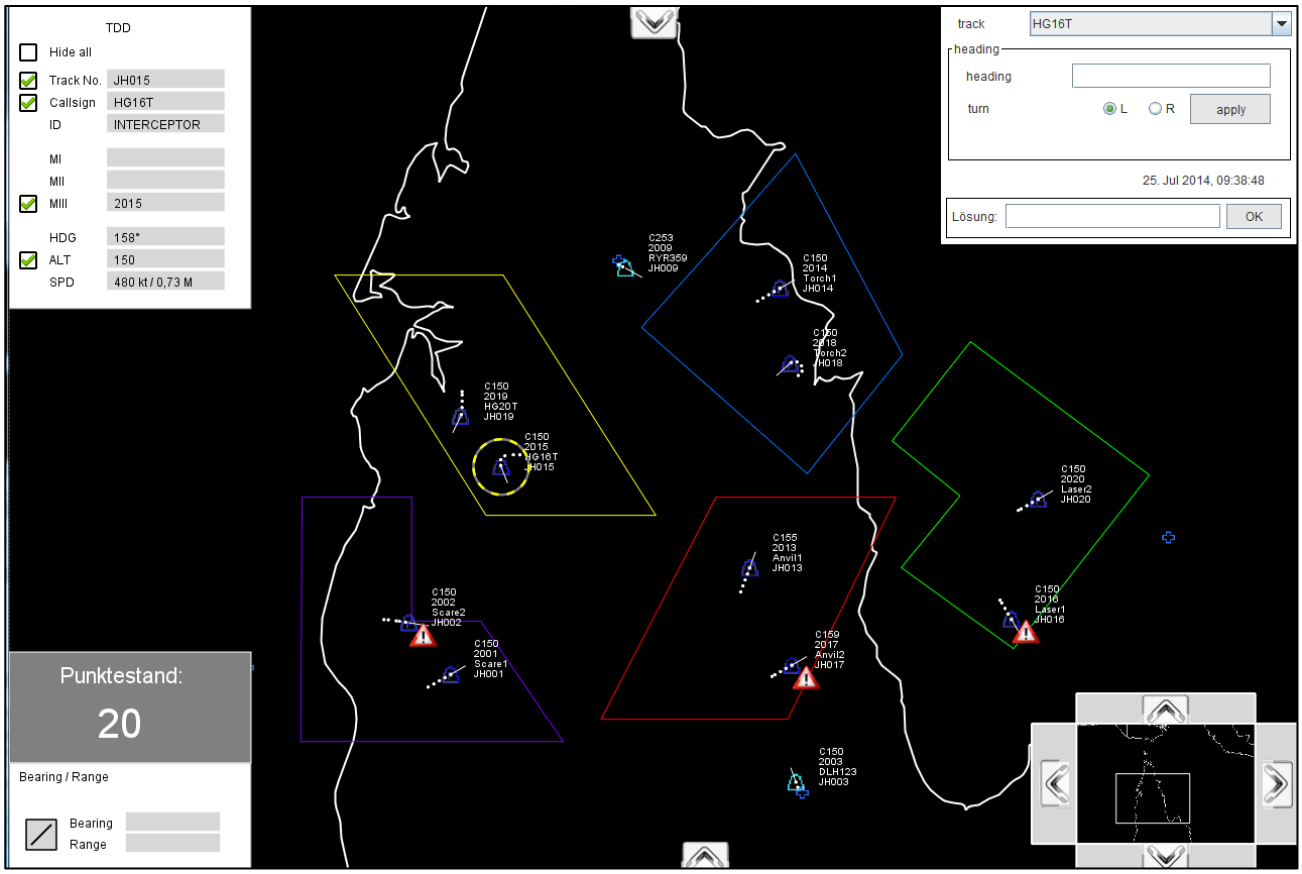


Abbildung 17. Verwendete Radarsimulation in Experiment 1

Die Lufträume sind als geometrische Formen farblich markiert. Pro Luftraum müssen zwei Luftkontakte – blau dargestellt – überwacht und gesteuert werden, wobei die Anzahl der zu überwachenden Lufträume je nach Versuchsbedingung variierte (vgl. Abschnitt 4.2.2). Kontakte in Türkis repräsentieren Fremdverkehr, der nicht steuerbar ist. Bei Annäherung eines Kontakts an eine Luftraumgrenze oder an einen anderen Kontakt müssen Anweisungen zur Kurskorrektur gegeben

werden, um den Konflikt zu vermeiden. Während dies im Realsystem per Funk erfolgt, wird im Experiment der entsprechende Kontakt per Mausklick ausgewählt und Kurs („heading“) und Drehrichtung („turn) werden in Formularfelder (in Abbildung 17 rechts oben) eingegeben. Bei Mausklick auf die „apply“-Schaltfläche setzt der ausgewählte Kontakt die Eingaben um und behält den eingegebenen Kurs mindestens 10 Sekunden bei, bevor wieder zufallsgesteuerte Richtungsänderungen einsetzen können. Zur Komplexitätsreduktion blieben im Experiment Flughöhe und Geschwindigkeit der Kontakte konstant und konnten nicht durch Eingaben des Probanden geändert werden.

Ein Nichteinhalten der geforderten Mindestabstände zwischen Kontakten (siehe Instruktion in Anhang A.3) sowie ein Überschreiten der Luftraumgrenze von mindestens einem Kontakt, stellt eine Regelverletzung dar, die zu einem Punktabzug führt (siehe Absatz zum Punktesystem). Drohende Regelverletzungen werden im Experimentalsystem durch ein Warnsymbol am Kontakt und einen auditiven Alarmton signalisiert. Die Alarmtöne für Kollisionswarnungen und Luftraumverletzungen sind unterschiedlich. Sie bleiben aktiv, bis eine Kurskorrektur vorgenommen wird.

Sekundäraufgabe: Rechenaufgaben lösen

Die Sekundäraufgabe dient dazu, weitere kognitive Anforderungen, die im operativen Betrieb auftreten, nachzubilden. Zum Einem steht der Operateur in Funkkontakt mit den Piloten, so dass nicht nur der visuelle sondern auch der auditive Wahrnehmungskanal beansprucht wird. Zum Anderen muss der Operateur Berechnungen von Flugparametern (z.B. Kurs, Höhe, Geschwindigkeit) vornehmen. Nach dem Multiple-Resource-Modell von Wickens (2002) werden somit räumliche und auch verbale kognitive Ressourcen beansprucht (vgl. Abschnitt 3.1.1). Um diese Anforderungen in der Experimentalaufgabe abzubilden, wird während der Laufzeit des Szenarios alle 45 Sekunden eine Rechenaufgabe auditiv auf Englisch präsentiert. Die Aufgabe ist nach folgendem Muster aufgebaut:

	dreistellige Zahl (auf Zehner gerundet)
+/-	zweistellige Zahl (auf Zehner gerundet)
+/-	einstellige Zahl

Ein Beispiel für eine Rechenaufgabe ist: 180+70-6. Die Rechenaufgabe ist an die Kursberechnung bei Manöverplanungen angelehnt. Die dreistellige Zahl repräsentiert den aktuellen Kurs des Kontakts, die zweistellige Zahl eine durchzuführende Kursänderung und die einstellige Zahl einen Korrekturwert zur Berücksichtigung der Flughöhe oder Windverhältnisse. Nach Ansage der Aufgabe hat der Proband 45 Sekunden Zeit, die Lösung in ein Textfeld einzugeben und die Eingabe zu bestätigen.

Punktesystem

Als Leistungsindex dient ein Punktesystem, das die fehlerfreie Durchführung der Aufgabe belohnt und Regelverletzungen sanktioniert. Neue Punkte und Punktabzüge werden kurzzeitig auf dem Bildschirm eingeblendet. Der sich daraus ergebende momentane Punktestand ist dauerhaft in der linken unteren Ecke des Bildschirms eingeblendet (vgl. Abbildung 17). Um die Motivation zu fördern, wurden die Probanden in der Instruktion (Anhang A.3) darauf hingewiesen, dass ein

positiver Punktestand am Ende des Versuchs belohnt wird. Dies sollte die Anstrengungsbereitschaft erhöhen und damit die Entstehung von Zuständen hoher Beanspruchung begünstigen (vgl. Abschnitt 3.1.3).

Alle 10 Sekunden werden 5 Pluspunkte vergeben, sofern kein Alarm aktiv ist und somit keine Regelverletzung droht. Richtig beantwortete Rechenaufgaben werden zusätzlich mit 5 Pluspunkten belohnt. Bei Regelverstößen werden Minuspunkte vergeben. Die Höhe der Minuspunkte spiegelt die Schwere des Verstoßes in der Realität wider. Da (Beinahe-)Kollisionen in der Realität am gravierendsten sind, werden sie mit 20 Minuspunkten bewertet. Darauf folgen Verstöße gegen die Luftraumgrenzen (10 Minuspunkte). Für falsche oder ausgelassene Berechnungen der Sekundäraufgabe fallen 5 Minuspunkte an. Nach einer festgelegten Zeit kommt es bei Nichtbehebung von Abstandsverletzungen (Kollision/Luftraumgrenze) zu erneuten Punktabzügen.

4.2.2 Versuchsdesign

In der experimentellen Untersuchung wurden drei Merkmale der Anforderungssituation (*Anzahl Areas, Kooperativität* und *Lärm)* systematisch variiert, um Veränderungen bei den Nutzerzustandsdimensionen Beanspruchung, Frustration und Aufmerksamkeit hervorzurufen. Es resultiert somit ein dreifaktorielles Versuchsdesign, in dem die drei Merkmale Faktoren mit jeweils zwei Faktorstufen darstellen. Die drei Faktoren und der Versuchsplan werden nachfolgend näher erläutert.

1. Faktor: Anzahl Areas (Variation der Belastung/Beanspruchung)

Der erste Faktor bezieht sich auf die Anzahl der zu überwachenden Lufträume (im Folgenden *Areas* genannt). Eine höhere Anzahl an Areas erhöht das Aufgabenvolumen, was dem CTL-Modell von Neerincx (2003) zufolge zu einer höheren Belastung führt und damit die mentale Beanspruchung des Versuchsteilnehmers erhöhen soll (vgl. Abschnitt 3.1.1). In der Konfiguration mit niedriger Belastung müssen *zwei Areas* und in der Konfiguration mit hoher Belastung *fünf Areas* überwacht werden.

2. Faktor: Kooperativität (Variation des Frustrationsniveaus)

Bei der Frustration handelt es sich um eine Emotion mit stark negativer Valenz, die sich negativ auf die Leistungsbereitschaft auswirken kann (vgl. Abschnitt 3.1.2). Studien weisen darauf hin, dass Frustration bei Computerspielen experimentell erzeugt werden kann, indem das System Tastatureingaben zur Steuerung von Objekten ignoriert. (vgl. Reuderink, Nijholt & Poel, 2009; Diener & Oertel, 2006). In der Studie von Reuderink et al. (2009) betraf dies 15% der Eingaben bei einem Pacman-Spiel, in der Studie von Diener & Oertel (2006) 20% der Eingaben bei einem Tetris-Spiel. Analog dazu wurde die Frustration in dem vorliegenden Experiment über den Grad der Kooperativität der Kontakte moduliert: In der Konfiguration mit *hoher Kooperativität* reagieren die Kontakte zu 100% auf Richtungsanweisungen. In der Konfiguration mit *geringer Kooperativität* besteht nur eine Wahrscheinlichkeit von 70%, dass ein Kontakt auf eine Anweisung reagiert. Die Wahrscheinlichkeit, dass der Kontakt die Anweisung ignoriert und zufallsgesteuert weiterfliegt, beträgt somit 30%. Sie liegt damit etwas höher als in den genannten Studien, da sich in Vortests zeigte, dass auf diese Weise Zustände hoher Frustration verlässlicher erzielt werden können.

3. Faktor Lärm (Variation der Aufmerksamkeit)

Lärm ist ein auditiver Stressor, der Befunden von Hockey (1970, 1986) zufolge das Arousal erhöht. Dies soll zu einer höheren Selektivität der Aufmerksamkeit führen, die sich negativ auf die Leistung auswirkt, wenn dadurch wichtige Aufgaben übersehen werden (vgl. Abschnitte 3.1.5 und 3.3.3). Ebenfalls stellten Trimmel & Poelzl (2006) fest, dass Hintergrundgeräusche die Leistung in Bezug auf räumliche Aufmerksamkeit verschlechtern. Zur Variation der Aufmerksamkeit wird zwischen den Konfigurationen *ohne Lärm* und *mit Lärm* unterschieden. In der Konfiguration mit Lärm hörten die Teilnehmer über Kopfhörer ein lautes konstantes Maschinengeräusch. Die Rechenaufgaben wurden ebenfalls auditiv dargeboten. Grundsätzlich war es allerdings trotz des Lärms möglich, die Aufgaben zu verstehen.

Versuchsplan

Der Versuchsplan mit den drei Faktoren *Anzahl Areas*, *Kooperativität der Kontakte* und *Lärm* und ihren jeweils zwei Faktorstufen ist in Abbildung 18 veranschaulicht. Um sowohl die Haupteffekte in Bezug auf die drei Faktoren als auch ihre Wechselwirkungen statistisch testen zu können, wurden sämtliche Faktorstufen systematisch miteinander kombiniert. Aus dem 2x2x2-faktoriellen Design ergeben sich somit acht Versuchsbedingungen (vgl. graue Nummerierung in Abbildung 18). Der Versuch ist als Messwiederholungsdesign konzipiert. Somit absolvierte jeder Versuchsteilnehmer nacheinander alle Bedingungen (8 Testdurchläufe). Um Müdigkeitseffekten vorzubeugen, wurden die Testdurchläufe auf zwei Versuchssitzungen (A und B) aufgeteilt, die an unterschiedlichen Tagen stattfanden und sich in Hinblick auf den Faktor *Lärm* unterschieden (vgl. Abbildung 18).

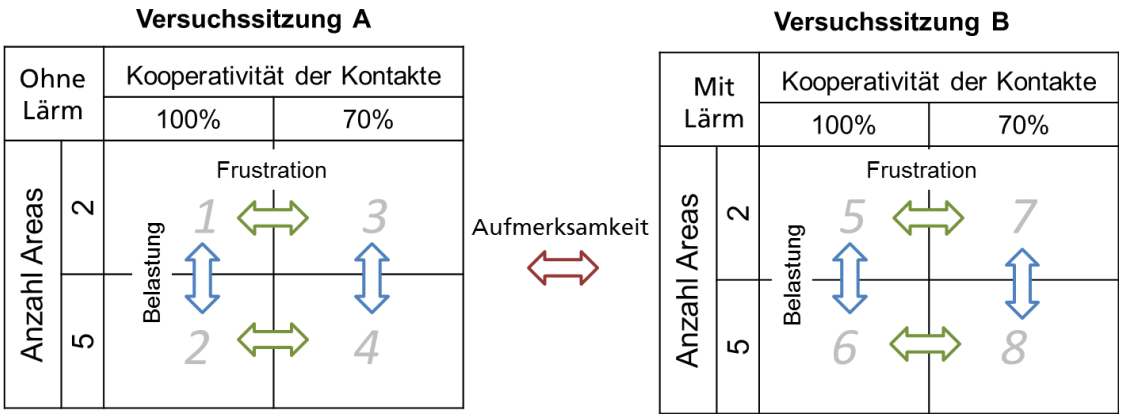


Abbildung 18. Versuchsdesign (Experiment 1)

Zur Kontrolle von Reihenfolge- und Übungseffekten wurde die Abfolge der Versuchsbedingungen innerhalb der Stichprobe variiert. Tabelle 16 zeigt einen Ausschnitt aus dem Testplan, in dem die Abfolge der Versuchsbedingungen für die ersten drei Versuchspersonen (VP) aufgeführt ist (der vollständige Testplan ist Anhang A.1 zu entnehmen). Die VP 1-6 absolvierten Versuchssitzung A (Tests 1-4) mit der Konfiguration ohne Lärm zuerst und Versuchssitzung B (Tests 5-8) mit der Konfiguration mit Lärm als Zweites. Bei den Versuchspersonen 7-12 war die Reihenfolge umgekehrt. Um Reihenfolgeeffekte innerhalb der Sitzungen A und B zu vermeiden, wurde die Reihenfolge der vier Versuchsbedingungen entsprechend einem Lateinischen Quadrat rotiert. Als Beispiel erhielt VP1 die Konfiguration mit 2 Areas und 100% Kooperativität im ersten Test, VP2 im zweiten und VP3 erhielt sie im dritten Test.

Tabelle 16. Testabfolge pro Versuchsperson in Experiment 1 (Ausschnitt)

		Versuchssitzung A				Versuchssitzung B			
VP	Faktor	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
1	Areas	2	5	2	5	2	5	2	5
	Koop.	100%	100%	70%	70%	100%	100%	70%	70%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
2	Areas	5	2	5	2	5	2	5	2
	Koop.	70%	100%	100%	70%	70%	100%	100%	70%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
3	Areas	2	5	2	5	2	5	2	5
	Koop.	70%	70%	100%	100%	70%	70%	100%	100%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja

4.2.3 Unabhängige und abhängige Variablen

Die in der experimentellen Untersuchung erfassten unabhängigen und abhängigen Variablen beziehen sich auf die im vereinfachten Nutzerzustandsmodell (vgl. Abbildung 16, S. 70) dargestellten Komponenten. Abbildung 19 führt die erfassten Variablen pro Komponente auf. Zu den Merkmalen der Anforderungssituation zählen die in Abschnitt 4.2.2 beschriebenen drei Faktoren Anzahl Areas, Kooperativität der Kontakte und Lärm. Zusammen mit den individuellen Faktoren stellen sie die unabhängigen Variablen (UV) dar, die sich auf den Nutzerzustand auswirken. Als abhängige Variablen (AV) fungieren die Maße, die durch den Nutzerzustand beeinflusst werden: Dies sind Maße zu Leistung, Physiologie und Verhalten sowie subjektive Maße (siehe nähere Beschreibungen in den Abschnitten 4.2.5 und 4.2.6). Die Buchstaben in Klammern (B, F, A) in Abbildung 19 zeigen an, für welche der experimentell modulierten Nutzerzustands-dimensionen die in der Untersuchung erfassten Variablen als Einflussfaktoren und Indikatoren herangezogen und untersucht werden sollen (B=Beanspruchung, F=Frustration, A=Aufmerksamkeit). In den folgenden Abschnitten werden diese Variablen und ihre Erfassungsmethoden detaillierter vorgestellt.

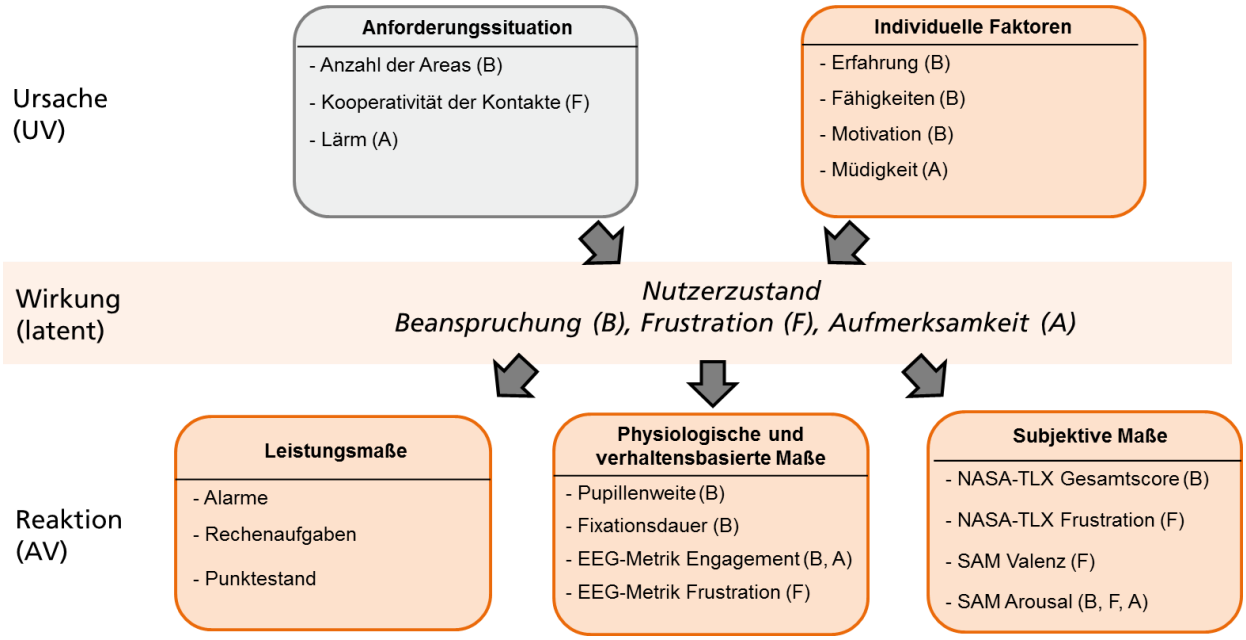


Abbildung 19. Übersicht über die in Experiment 1 erfassten unabhängigen und abhängigen Variablen

4.2.4 Individuelle Faktoren[3]

Individuelle Faktoren stellen in der Untersuchung unabhängige Variablen dar, deren Einfluss auf die Leistung und den Nutzerzustand untersucht werden sollte. Die Versuchspersonen wurden nicht vorab nach bestimmten individuellen Merkmalen ausgewählt. Es wurden jedoch einige als relevant erachtete individuelle Merkmale zu Beginn des ersten und zweiten Durchgangs erhoben. Dazu zählen neben der *Erfahrung* und den *Fähigkeiten* des Versuchsteilnehmers in diesem Fall auch die Nutzerzustandsdimensionen *Motivation* und *Müdigkeit*. Da diese Zustände in der Untersuchung nicht gezielt verändert wurden, werden sie als Aspekte des „Baselinezustands" betrachtet (vgl. Abschnitt 3.3.1). Die Erfassung der individuellen Faktoren erfolgte über die im Folgenden beschriebenen Tests und Fragebögen. Sie wurden über ein am Fraunhofer FKIE entwickeltes Softwaretool auf dem Computer dargeboten.

Erfahrung

Da es sich bei den Teilnehmern um Novizen handelte (vgl. Abschnitt 4.2.8), haben sie prinzipiell keine Vorerfahrungen mit der Experimentalaufgabe. Dennoch ist anzunehmen, dass Erfahrung im Umgang mit computerbasierten Simulations- und Strategiespielen die Bearbeitung der Aufgabe erleichtern kann. Die Teilnehmer wurden daher vor der Untersuchung danach befragt, wie gut sie sich mit computerbasierten Simulations- und Strategiespielen auskennen. Die Teilnehmer konnten dabei zwischen *gar nicht*, *ein wenig*, *gut* und *sehr gut* wählen.

Fähigkeiten

Kognitive Fähigkeiten, die für das Steuern der Luftkontakte in der Primäraufgabe und das Lösen der Rechenaufgaben in der Sekundäraufgabe erforderlich sind, wurden über zwei selbst entwickelte Leistungstests erfasst. Sie sind konzeptionell an kognitive Tests von etablierten Testverfahren angelehnt, wie *ANAM* (Automated Neuropsychological Assessment Metrics, Reeves et al., 2007) oder *CogGauge* (A Cognitive Assessment Tool for Spaceflight, Johnston, Carpenter & Hale, 2011), beziehen sich aber auf den Kontext und die spezifischen Anforderungen der Experimentalaufgabe.

Der „Links-Rechts-Test"

Bei der Steuerung der Kontakte auf dem Radarbildschirm liegt die Schwierigkeit darin, dass Kurs und Drehrichtung (links oder rechts) jeweils aus der Perspektive des zu steuernden Kontakts angegeben werden müssen und somit von dessen Ausrichtung abhängen. Neben der Fähigkeit zur Links-Rechts-Unterscheidung erfordert dies auch die Fähigkeit zur räumlichen Orientierung. Der selbstentwickelte Links-Rechts-Test (kurz: L-R-Test) prüft diese Fähigkeiten ab. Konzeptionell ist dieser Test mit bestehenden Tests zur räumlichen Orientierung vergleichbar (z.B. *Manikin* aus der Testbatterie ANAM, Reeves et al., 2007). Wie in Abbildung 20 zu sehen ist, werden jeweils ein Luftkontakt (Anvil 1) und ein Flughafen (ETNZ) dargestellt. Aufgabe bei diesem Test ist es, so

[3] Die Bezeichnung als „Faktor" geht auf die Analyse der Einflussfaktoren in Kapitel 3 zurück und bezeichnet keinen experimentell manipulierten Faktor. Gleichbedeutend mit Faktor ist in diesem Fall die Bezeichnung Merkmal.

schnell wie möglich anzugeben, in welche Richtung (links, also gegen den Uhrzeigersinn oder rechts, im Uhrzeigersinn) sich der Luftkontakt drehen muss, um den Flughafen auf kürzestem Weg zu erreichen. Die Drehrichtung muss – analog zur Experimentalaufgabe – mit der Maus ausgewählt werden. Der Test beinhaltet 32 Aufgaben, bei denen sich die Ausrichtung des Luftkontakts und die Position des Flughafens jeweils unterscheiden.

Abbildung 20. Aufgabe aus dem „Links-Rechts-Test“

Der „Rechentest“

Der Rechentest wurde entwickelt, um die Fähigkeiten im Rechnen erfassen zu können, die für die Bearbeitung der Sekundäraufgabe benötigt werden. Er hat Ähnlichkeit mit dem Test *Mental Processing* aus der Testbatterie ANAM (Reeves, Winter, Bleiberg, & Kane, 2007) sowie *Asteroid Sling* von CogGauge (Johnston, Carpenter, & Hale, 2011), wobei die im Rechentest gestellten Aufgaben der Form nach exakt den Rechenaufgaben in der Experimentalaufgabe entsprechen (vgl. Abschnitt 4.2.1). Der Test beinhaltet 25 Rechenaufgaben, die hintereinander durch Eingabe der Lösung in ein Textfeld, so schnell und so korrekt wie möglich, beantwortet werden müssen.

Auswertung der Tests

Bei kognitiven Tests werden die Fähigkeiten üblicherweise anhand der Genauigkeit (Anzahl korrekter Aufgaben) und der Zeitdauer der Aufgabenbearbeitung bewertet. Dabei muss beachtet werden, dass sich Teilnehmer darin unterscheiden können, ob sie mehr Wert auf die Geschwindigkeit oder die Genauigkeit legen (Dickman & Meyer, 1988). Es zeigte sich, dass eine höhere Genauigkeit häufig zu Lasten der Geschwindigkeit geht und eine höhere Geschwindigkeit zu Lasten der Genauigkeit (sog. „Speed-Accuracy-Tradeoff“, vgl. Wickelgren, 1977). In der vorliegenden Arbeit wurden daher beide Indikatoren zur Bewertung der Fähigkeiten herangezogen, wobei diese zu einem einzelnen Bewertungsmaß fusioniert wurden. Ein weit verbreitetes Verfahren hierfür ist die Berechnung des *IES (Inverse Efficiency Score)* nach Townsend & Ashby (1978). Die Reaktionszeit bei den korrekt beantworteten Aufgaben wird hierbei durch den prozentualen Anteil

korrekt beantworteter Aufgaben geteilt. Hohe Werte weisen dabei allerdings auf eine schlechtere Leistung hin, was die Interpretation der Ergebnisse möglicherweise erschwert. Daher wurde ein reziprokes Maß verwendet, das von Woltz & Was (2006) entwickelt und von Vandierendonck (2017) als *RCS (Rate Correct Score)* bezeichnet wird. Die Berechnung des RCS ist in Gleichung 1 wiedergegeben. Demnach ergibt sich der RCS aus der Anzahl korrekter Antworten im Test (in Gleichung 1 das *c*), die durch die Summe der Reaktionszeiten (RT) bei allen Aufgaben geteilt wird.

$$RCS = \frac{c}{\sum RT}$$

Gleichung 1. Berechnung des Rate Correct Score (RCS)

Das Ergebnis ist ein Wert, der die durchschnittliche Anzahl korrekter Antworten pro Zeiteinheit anzeigt. Höhere Werte gehen somit mit einer besseren Leistung im Test einher. Bei einem Vergleich der Leistungswerte von RCS und IES in der vorliegenden Untersuchung zeigt sich, dass die Maße mit $r > -.95$ sehr hoch miteinander korrelieren. Beide Verfahren können daher für diesen Kontext als weitgehend äquivalent eingeschätzt werden.

Schlafbezogene Müdigkeit

Schlafbezogene Müdigkeit kann die Leistungsfähigkeit in starkem Maße beeinträchtigen (vgl. Abschnitt 3.1.4). Da auf das Schlafverhalten und die Müdigkeit der Teilnehmer kein Einfluss genommen werden konnte, wurde die schlafbezogene Müdigkeit vorab über eine Ratingskala erfasst. Hierfür wurde die *Stanford Sleepiness Scale* (Hoddes et al., 1973) verwendet. Sie besteht aus einer Skala mit sieben Abstufungen, die den momentanen Zustand der Schläfrigkeit/ schlafbezogenen Müdigkeit beschreiben soll. In der Untersuchung wurde eine deutschsprachige Übersetzung der Stanford Sleepiness Scale nach Mieg (2006) eingesetzt (siehe Anhang A.6).

Motivation

Hinsichtlich der Motivation ist davon auszugehen, dass sich motivierte Teilnehmer im Vergleich zu weniger motivierten Teilnehmern mehr anstrengen und somit eine bessere Leistung bei der Experimentalaufgabe erzielen (vgl. Abschnitt 3.1.3). Es gibt nur wenige Fragebögen, die in Betracht kommen, die Motivation im vorliegenden Kontext zu erfassen. Dem Untersuchungszweck am nächsten kommt der *FAM* (Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen; Rheinberg, Vollmeyer & Burns, 2001). Er ist auf die Erfassung aktueller Motivation ausgerichtet, bezieht sich dabei allerdings auf den Lernkontext. Die *PRS-Skala* von Adair (1970; in der deutschen Fassung von Timaeus, Lück, Klandt & Schanderwitz, 1977) wird von Timaeus et al. (1977) zur Kontrolle der Motivation von Versuchspersonen in Experimenten vorgeschlagen. Mit der PRS-Skala werden jedoch vorwiegend die allgemeinen Einstellungen zu Experimenten und weniger die aktuelle Motivation erfasst. Beide Fragebögen sind mit 18 Items beim FAM und 39 Items beim PRS außerdem recht lang. Es wurde daher als zweckmäßiger erachtet, die Motivation über eine selbst generierte Frage zu erfassen. Bei dieser sollten die Teilnehmer angeben, wie hoch sie ihre Motivation einschätzen, bei der Untersuchung eine gute Leistung zu erzielen. Die Bewertung erfolgte auf einer fünfstufigen Likert-Skala mit den Polen 1 „sehr hoch“ und 5 „sehr gering“.

4.2.5 Physiologische und verhaltensbasierte Maße

In der Bewertung der Erfassungsmethoden in Abschnitt 2.3 hatte sich herausgestellt, dass okulomotorische Maße sowie Maße der Hirnaktivität gute Voraussetzungen für eine Echtzeitbewertung des multidimensionalen Nutzerzustands bieten: Sie können kontinuierlich über einen Eyetracker und ein EEG aufgezeichnet werden, die Daten sind in nahezu Echtzeit auswertbar und sie können als Indikatoren für verschiedene Nutzerzustände herangezogen werden. In der experimentellen Untersuchung wurde nun geprüft, inwiefern Metriken eines Eyetrackers und eines EEG in der Lage sind, Veränderungen hinsichtlich der Nutzerzustände Beanspruchung, Frustration und Aufmerksamkeit valide zu erfassen. Als Eyetracker diente der Remote-Eyetracker *Tobii X120* (vgl. Abbildung 21 links). Als EEG-Sensor wurde das Headset *Emotiv EPOC™* verwendet (vgl. Abbildung 21 rechts). Im Folgenden werden die Sensoren sowie die Metriken, die für die Nutzerzustandserfassung herangezogen wurden, näher vorgestellt.

Abbildung 21. Eyetracker Tobii X120 (links) und das EEG Headset Emotiv EPOC™ (rechts)

Metriken des Tobii X120 Eyetrackers

Der Eyetracker Tobii X120 ist ein Remote-Eyetracker, der unterhalb des Bildschirms platziert wird. Das Gerät zeichnet Blickbewegungen und Pupillenmaße mit einer Frequenz von 120 Hz auf. Die Auswertung erfolgte mit der herstellereigenen Software *Tobii Studio™ 2.2*. Über ein mitgeliefertes SDK (Software Development Kit) ist auch ein Echtzeit-Zugriff auf die aufgezeichneten Parameter möglich.

Als Diagnosemaße wurden die durch Tobii Studio™ zur Verfügung gestellten Parameter *Pupillenweite* und *Fixationsdauer* herangezogen. Die Pupillenweite ist ein globales Beanspruchungsmaß, wobei hohe mentale Beanspruchung mit einer höheren Pupillenweite einhergeht (vgl. Abschnitt 2.3.3). In Tobii Studio™ wird die Pupillenweite für das linke und das rechte Auge in mm ausgegeben. Die Fixationsdauer ist von der Aufgabenart abhängig. Studien, welche die Fixationsdauer zur Erfassung des mentalen Zustands bei visuell beanspruchenden Aufgaben verwendeten, weisen darauf hin, dass sich die Fixationsdauer mit zunehmender visueller Beanspruchung verringert (z.B. DeRivecourt et al., 2008) und dass sie negativ mit der Leistung zusammenhängt (Van Orden et al., 2000). Zur Berechnung von Fixationen wird von Tobii Studio™ ein von Olsson (2007) vorgeschlagener Filteralgorithmus angewendet, der sich ähnlich wie ein I-VT Filter verhält (vgl. Tobii, 2010; Olsen, 2012). Da sich der Blickpunkt auch bei Fixationen geringfügig bewegt, analysiert der Algorithmus anhand der Geschwindigkeit der Blickbewegung,

ob diese noch Teil einer Fixation ist oder als Sakkade zu werten ist. Die Mindestdauer, die eine Fixation aufweisen muss, um als solche detektiert zu werden, kann vom Nutzer selbst festgelegt werden. In der experimentellen Untersuchung wurde die von Tobii (2012) vorgeschlagene Mindestdauer von 60 ms verwendet.

Metriken des Emotiv EPOC^{TM} EEG

Das EEG-System Emotiv EPOC^{TM} ist ein kabelloses Multikanal-EEG-Headset, das die elektrische Hirnaktivität im Frequenzbereich von 0,2-43 Hz aufzeichnet. Es besteht aus 14 Elektroden und zwei Referenzelektroden, die an Plastikarmen befestigt sind. Die Elektroden werden gemäß dem internationalen 10/20 System an standardisierten Positionen auf der Kopfhaut platziert (vgl. Jurcak, Tsuzuki, & Dan, 2007). Im Gegensatz zu vielen anderen EEG-Systemen wird kein leitendes Gel benötigt, um den Kontakt zwischen der Kopfhaut und den Elektroden herzustellen. Stattdessen werden die an den Elektroden angebrachten Filzpads mit Kochsalzlösung befeuchtet.

Emotiv EPOC^{TM} verfügt über eine Klassifikationssoftware (*Affectiv^{TM}Suite*), die über ein SDK Echtzeit-Auswertungen für verschiedene mentale Zustände zur Verfügung stellt. Dabei handelt es sich um die Metriken *Engagement*, *Frustration*, *Excitement* und *Meditation*. Die Metriken werden auf einer Skala von 0 bis 1 ausgegeben, wobei höhere Werte resultieren, je höher der Ausprägungsgrad des zu messenden Zustands ist (Emotiv, 2013). Für die Nutzung ist weder ein Training noch eine Kalibrierung notwendig. Stattdessen passen sich die Metriken von selbst durch sogenanntes „self-scaling“ (Emotiv Nutzerforum, 2010a) an den jeweiligen Nutzer an. Dies führt allerdings dazu, dass die Werte zwischen den Nutzern nicht vergleichbar sind. In der vorliegenden Untersuchung wurden für die Auswertungen auf Gruppenebene daher nicht die absoluten sondern die an einem individuellen Referenzwert relativierten Werte der EEG-Metriken verwendet (vgl. Abschnitt 4.2.9).

Die Algorithmen, die den Metriken zu Grunde liegen, wurden von Emotiv selbst entwickelt und sind der Öffentlichkeit nicht zugänglich. Auch die Beschreibungen zu den Metriken sind im Manual (Emotiv, 2013) sehr knapp gehalten. Daher wurde auch das Nutzerforum von Emotiv als Informationsquelle hinzugezogen. Ein Administrator berichtet darin, dass die Entwicklung der Metriken auf Experimenten basierte, in denen unterschiedliche Stimuli verwendet wurden, um die interessierenden Zustände zu erzeugen. Die dabei aufgezeichneten Daten des EEG wurden mit Expertenbewertungen und anderen physiologischen Indikatoren (z.B. Herzrate, Atmungsrate, Blutdruck, elektrodermale Aktivität) verglichen und statistisch ausgewertet, um die besten Indikatorvariablen für jeden Zustand zu identifizieren (Emotiv Nutzerforum, 2010b). Beschreibungen der einzelnen Metriken aus dem Manual und dem Nutzerforum sind in Tabelle 17 zusammengefasst.

Die Beschreibung der Metrik Engagement (im Folgenden *EEG-Engagement* genannt) weist darauf hin, dass sie mit den im Experiment modulierten Nutzerzuständen mentale Beanspruchung und Aufmerksamkeit korrespondiert und bei hoher Beanspruchung und hoher Fokussierung der Aufmerksamkeit zunimmt. Für die Entwicklung der Metrik Frustration (im Folgenden *EEG-Frustration* genannt) wurde Frustration auf ähnliche Weise provoziert wie im Experiment bei der Konfiguration mit geringer Kooperativität der Kontakte (vgl. Tabelle 17). Es ist somit anzunehmen, dass in Bedingungen mit geringer Kooperativität der Kontakte höhere Frustrationswerte erzielt

werden als in Bedingungen mit hoher Kooperativität. Die Metriken Excitement und Meditation korrespondieren nicht direkt mit Nutzerzuständen, die im Experiment gezielt moduliert wurden. Auf eine inferenzstatistische Auswertung dieser Metriken wurde daher verzichtet.

Tabelle 17. Beschreibung der Metriken von Emotiv EPOC™

Emotiv-Metrik	Beschreibung	Korrespondierender Nutzerzustand
Engagement	Im Manual charakterisiert als „bewusstes Richten der Aufmerksamkeit auf aufgabenrelevante Stimuli“. Je größer die Aufmerksamkeit, Fokussierung und kognitive Beanspruchung, desto größer der Ausgabewert. Verwandte Zustände: Aufmerksamkeit, Vigilanz, Konzentration, Stimulierung und Interesse. (Emotiv, 2013, S. 31)	Mentale Beanspruchung, Aufmerksamkeit
Frustration	Keine Beschreibung im Manual. Emotiv Nutzerforum (2010b): Für die Entwicklung wurden Videospiele verwendet, die so gestaltet waren, dass sie den Nutzer frustrierten (z.B.Steuerung einer Videofigur wurde durch zufällige Richtungswechsel erschwert).	Frustration als emotionaler Zustand mit negativer Valenz
Excitement	Im Manual beschrieben als Gefühl physiologischer Erregung mit positiver Valenz, das durch eine Aktivierung des sympathischen Nervensystems charakterisiert ist. Verwandte Zustände: Aufregung, Erregung, Nervosität (Emotiv, 2013, S. 31).	Emotionaler Zustand mit positiver Valenz
Meditation	Keine Beschreibung im Manual. Emotiv Nutzerforum (2011): Korrespondiert mit ruhigen entspannten Zuständen, bleibt aber niedrig, solange kein echter meditativer Zustand erreicht ist.	Keine Entsprechung

4.2.6 Subjektive Maße und Leistungsmaße

Im Folgenden wird dargelegt, welche Verfahren und Maße zur subjektiven Bewertung der Beanspruchung und des emotionalen Zustands sowie zur Bewertung der Leistung herangezogen wurden. In Hinblick auf die Aufmerksamkeit ist kein geeigneter Fragebogen bekannt, der räumliche Aufmerksamkeit erfasst. Zudem ist anzunehmen, dass diese den Nutzern nicht bewusst ist. Insofern wird auf eine subjektive Bewertung der Aufmerksamkeit verzichtet.

Erfassung der Beanspruchung

Zur Erfassung subjektiver Beanspruchung wird der Fragebogen *NASA-TLX* von Hart & Staveland (1988) herangezogen. Dies ist ein multidimensionaler Fragebogen, der die Beanspruchung über sechs Subskalen (*geistige Anforderungen, körperliche Anforderungen, zeitliche Anforderungen, Leistung, Anstrengung, Frustration*) erfasst. Die Bewertungen auf den einzelnen Skalen können anschließend zu einem allgemeinen Gesamtscore zusammengefasst werden. Auf eine bei Hart & Staveland (1988) vorgesehene Gewichtung der Skalen wird bezugnehmend auf Byers, Bittner, & Hill (1989), Nygren (1991) und Pfendler, Pitrella, & Wiegand (1995) verzichtet.

Die Bewertungen wurden, wie bereits in vorangegangenen Untersuchungen (Schwarz & Witt, 2011; Witt, Özyurt, Schwarz, Döring, & Dörfel, 2012; Schwarz, Fuchs, Becker, & Kaster, 2013), mit der *KU-Skala* (Kategorienunterteilungsskala, Heller, 1982) vorgenommen. Dabei handelt es sich um eine 15-stufige Skala, die in fünf verbale Oberkategorien mit je drei Abstufungen unterteilt ist. Die Skala wurde so konzipiert, dass sie differenzierte und reliable Beurteilungen von Empfindungen

ermöglicht (Heller, 1982; Keilhacker, 2013). Der Fragebogen mit den sechs NASA-TLX Skalen ist in Anhang A.7 aufgeführt.

Erfassung des emotionalen Zustands

Zur Erfassung des emotionalen Zustands wurde der Fragebogen *SAM* (*Self Assessment Manikin)* eingesetzt (Bradley & Lang, 1994; vgl. Anhang A.5). Dies ist ein non-verbales Verfahren, bei dem der emotionale Zustand anhand von Piktogrammen (sogenannten Manikins) beurteilt wird. Die Beurteilung bezieht sich auf die drei Dimensionen *Valenz, Arousal* und *Dominanz*, die nach Bradley & Lang (1994) das affektive Erleben grundlegend bestimmen. Da Frustration dem Modell von Russell (1980) zufolge durch negative Valenz und hohes Arousal charakterisiert ist (vgl. Abschnitt 3.1.2), sind für die vorliegende Untersuchung insbesondere die Dimensionen *Valenz* und *Arousal* von Interesse. Die Dimensionen werden auf einer 9-stufigen Skala bewertet, wobei jede zweite Stufe durch ein Piktogramm veranschaulicht wird. Die Skala reicht für die Valenz von negativ (1) bis positiv (9) und für das Arousal von ruhig (1) bis erregt (9).

Erfassung der Leistung

Da das Auftreten von Alarmen auf Regelverletzungen hinweist, wurde die Dauer von Alarmen als Leistungsmaß für die Primäraufgabe herangezogen. In Bezug auf die Sekundäraufgabe wurde die Leistung über die Anzahl an ausgelassenenen Rechenaufgaben bestimmt. Der Punktestand bei Szenarioende wurde herangezogen, um die Gesamtleistung zu beurteilen.

4.2.7 Hypothesen

Entsprechend dem generischen Nutzerzustandsmodell (siehe Abbildung 15) und auf Basis der Erkenntnisse aus den theoretischen Analysen in Kapitel 2 und 3 wurde angenommen, dass zwischen den betrachteten unabhängigen und abhängigen Variablen Zusammenhänge und Abhängigkeiten bestehen. Diese wurden über drei übergeordnete Forschungshypothesen untersucht:

H1: Der diagnostizierte Nutzerzustand verändert sich in Abhängigkeit von der Anforderungssituation.

H2: Individuelle Faktoren beeinflussen den diagnostizierten Nutzerzustand und die Leistung.

H3: Leistungsmaße, physiologische/verhaltensbasierte Maße und subjektive Maße korrelieren miteinander.

Die Hypothesen 1 und 2 beziehen sich auf das Forschungsziel, den Einfluss von Merkmalen der Anforderungssituation und individuellen Faktoren auf den Nutzerzustand zu untersuchen. Durch Hypothese 3 soll in Hinblick auf das zweite Forschungsziel die Eignung physiologischer und verhaltensbasierter Maße als Indikatoren des Nutzerzustands überprüft werden (vgl. Abschnitt 4.1). Jede der drei Hypothesen unterteilt sich in drei Subhypothesen (a-c). Abbildung 22 visualisiert diese Hypothesen auf Grundlage des in Abschnitt 4.2.3 beschriebenen Modells zu unabhängigen und abhängigen Variablen (vgl. Abbildung 19). Einfache Pfeile weisen auf eine Ursache-Wirkung-Relation, Doppelpfeile auf eine wechselseitige Beeinflussung (Korrelation) hin.

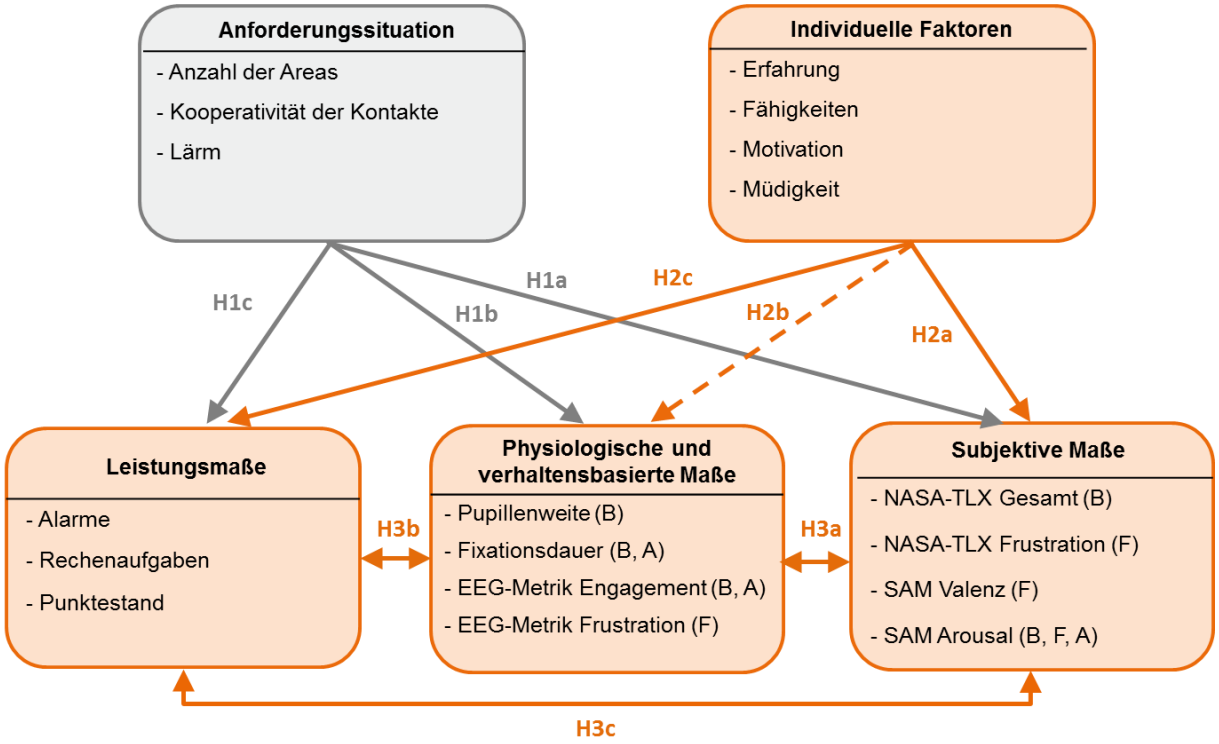


Abbildung 22. Hypothetisierte Abhängigkeiten und Zusammenhänge zwischen den Variablen (Experiment 1)

Die Subhypothesen beinhalten wiederum mehrere statistische Hypothesen, welche sich auf die Zusammenhänge zwischen den jeweiligen Einzelmaßen beziehen (in Abbildung 22 nicht explizit dargestellt). Sie werden nachfolgend für jede Hypothese näher ausgeführt.

H1: Der diagnostizierte Nutzerzustand verändert sich in Abhängigkeit der Anforderungssituation.

Hypothese 1 nimmt an, dass die drei Merkmale der Anforderungssituation, die als Faktoren experimentell variiert wurden, Nutzerzustandsveränderungen hevorrufen und sich damit auf die verwendeten Diagnosemaße (Leistungsmaße, physiologische/verhaltensbasierte Maße, subjektive Maße) auswirken. Gemäß H1 sollten daher signifikante Mittelwertunterschiede zwischen den Faktorstufen der Faktoren *Areas*, *Kooperativität* und *Lärm* bestehen, und zwar in Bezug auf:

- Subjektive Maße (H1a),
- Physiologische und verhaltensbasierte Maße (H1b),
- Leistungsmaße (H1c).

Zur Untersuchung der Subhypothesen H1a-H1c wurden pro Methodenkategorie unterschiedliche Diagnosemaße herangezogen (siehe Abschnitte 4.2.5 und 4.2.6). Die statistischen Hypothesen zu den einzelnen Diagnosemaßen wurden aus Literaturbefunden abgeleitet (vgl. Abschnitte 4.2.5 und 4.2.6). Sie sind in Tabelle 18 zusammengefasst.

Tabelle 18. Angenommene Mittelwertunterschiede zwischen den Faktorstufen für die subjektiven Maße (H1a), die physiologischen und verhaltensbasierten Maße (H1b) und die Leistungsmaße (H1c) – Experiment 1

		Anzahl Areas		Kooperativität		Lärm	
		A1: 2 Areas	A2: 5 Areas	K1: 100%	K2: 70%	L1: Ohne Lärm	L2: Mit Lärm
H1a	NASA-TLX-Gesamtscore	$\mu_{A1} < \mu_{A2}$					
	NASA-TLX Frustration			$\mu_{K1} < \mu_{K2}$			
	SAM Valenz			$\mu_{K1} > \mu_{K2}$			
	SAM Arousal	$\mu_{A1} < \mu_{A2}$		$\mu_{K1} < \mu_{K2}$		$\mu_{L1} < \mu_{L2}$	
H1b	Pupillenweite	$\mu_{A1} < \mu_{A2}$					
	Fixationsdauer	$\mu_{A1} > \mu_{A2}$					
	EEG-Engagement	$\mu_{A1} < \mu_{A2}$				$\mu_{L1} < \mu_{L2}$	
	EEG-Frustration			$\mu_{K1} < \mu_{K2}$			
H1c	Dauer Alarme	$\mu_{A1} < \mu_{A2}$		$\mu_{K1} < \mu_{K2}$		$\mu_{L1} < \mu_{L2}$	
	Ausgelassene Rechenaufgaben	$\mu_{A1} < \mu_{A2}$		$\mu_{K1} < \mu_{K2}$		$\mu_{L1} < \mu_{L2}$	
	Punktestand	$\mu_{A1} > \mu_{A2}$		$\mu_{K1} > \mu_{K2}$		$\mu_{L1} > \mu_{L2}$	

Anmerkung zu Valenz: kleinere Werte signalisieren einen negativeren Gefühlszustand; Grau: es wurde keine Hypothese definiert.

H2: Individuelle Faktoren beeinflussen den diagnostizierten Nutzerzustand und die Leistung.

In Kapitel 3 wurde festgestellt, dass sich variable und überdauernde individuelle Faktoren auf den Nutzerzustand und die Leistung auswirken. In Hypothese 2 wird daher angenommen, dass die individuellen Faktoren *Fähigkeiten*, *Erfahrungsgrad*, *Müdigkeit* und *Motivation* interindividuelle Unterschiede in dem subjektiv bewerteten Nutzerzustand und der Leistung erklären können. Konkret werden (Kausal-)Zusammenhänge zwischen den individuellen Faktoren und folgenden abhängigen Variablen angenommen:

- dem *NASA-TLX-Gesamtscore* und der SAM-Dimension *Arousal* als subjektive Maße des Nutzerzustands (H2a),
- den Leistungsmaßen: *Auslassungen in den Rechenaufgaben, Dauer der Alarme* und *Punktestand* (H2c).

An dieser Stelle sei darauf hingewiesen, dass für die physiologischen und verhaltensbasierten Maße Effekte zwischen Personen aufgrund der vorgenommenen z-Standardisierung der Werte nicht untersucht werden können (vgl. Abschnitt 4.2.9). Hypothese H2b existiert daher nur theoretisch und wurde nicht statistisch getestet.

Die Untersuchung der Hypothesen H2a und H2c erfolgte – auch wenn es sich um Annahmen zu Kausalzusammenhängen handelt – korrelativ (siehe nähere Ausführungen hierzu in Abschnitt 4.2.10). Für die Hypothesentestung wurden die in Tabelle 19 aufgeführten gerichteten statistischen Hypothesen aufgestellt.

Tabelle 19. Angenommene Richtung der Korrelationen für die in den Hypothesen H2a und H2b erwarteten (Kausal-)Zusammenhänge (Experiment 1)

		L-R-Test	Rechnen	Erfahrung	Müdigkeit	Motivation
H2a	NASA-TLX	$\rho < 0$	$\rho < 0$	$\rho < 0$		$\rho > 0$
	SAM-Arousal				$\rho < 0$	$\rho > 0$
H2c	Dauer Alarme	$\rho < 0$	$\rho < 0$	$\rho < 0$	$\rho > 0$	$\rho < 0$
	Auslassungen Rechnen	$\rho < 0$	$\rho < 0$	$\rho < 0$	$\rho > 0$	$\rho < 0$
	Punktestand	$\rho > 0$	$\rho > 0$	$\rho > 0$	$\rho < 0$	$\rho > 0$

Grau: keine Hypothese definiert; die gleichen Annahmen gelten für den Regressionskoeffzienten β bei Berechnung bivariater Regressionen.

In Bezug auf die Hypothese H2a wird erwartet, dass gute Fähigkeiten in der Links-Rechts-Unterscheidung und im Rechnen sowie ein hoher Erfahrungsgrad dazu führen, dass die mentale Beanspruchung geringer ausgeprägt ist (vgl. Abschnitt 3.1.1). Diese Faktoren sollten somit negativ mit dem NASA-TLX-Gesamtscore korrelieren. Hohe Motivation sollte hingegen die Bereitschaft, sich anzustrengen, erhöhen (vgl. Abschnitt 3.1.3) und positiv mit dem NASA-TLX-Gesamtscore korrelieren. In Hinblick auf die SAM-Dimension Arousal wird vermutet, dass Motivation das Arousal erhöht, während Müdigkeit es verringert.

Hinsichtlich Hypothese H2c wird angenommen, dass die Fähigkeiten in der Links-Rechts-Unterscheidung und im Rechnen sowie der Erfahrungsgrad Leistungsunterschiede in der Experimentalaufgabe zwischen den Teilnehmern erklären können. Konkret wird angenommen, dass Personen mit einer besseren Leistung in den Vorabtests und Personen mit mehr Erfahrung in Computerspielen in den Versuchsdurchläufen eine geringere Dauer von Alarmen und weniger ausgelassene Rechenaufgaben aufweisen. In Hinblick auf das Leistungsmaß Punktestand wird hingegen ein positiver Zusammenhang erwartet. Hinsichtlich der Variablen Müdigkeit und Motivation wird auf Basis der theoretischen Erkenntnisse in Kapitel 3 angenommen, dass hohe Müdigkeit die Leistung verschlechtert, hohe Motivation die Leistung hingegen verbessert.

H3: Leistungsmaße, physiologische/verhaltensbasierte Maße und subjektive Maße korrelieren miteinander.

Mit Hypothese 3 soll die konvergente Validität der physiologischen und verhaltensbasierten Maße überprüft werden. Diese liegt dann vor, wenn sie mit Diagnosemaßen, die den gleichen Nutzerzustand erfassen, hoch korrelieren (Eid, 2014). Untersucht wurden die Korrelationen zwischen:

- subjektiven Maßen und physiologischen/verhaltensbasierten Maßen (H3a),
- Leistungsmaßen und physiologischen/verhaltensbasierten Maßen (H3b),
- sowie zu Vergleichszwecken auch zwischen subjektiven Maßen und Leistungsmaßen (H3c).

Für die jeweiligen Variablenpaare wurden anhand der literaturbasierten Befunde (vgl. Kapitel 2, Abschnitt 2.3) und den Variablenbeschreibungen in den Abschnitten 4.2.5 und 4.2.6 gerichtete Zusammenhangshypothesen formuliert, die in Tabelle 20 dargestellt sind. Um den Umfang der

Analysen zu beschränken, wurde bei den Leistungsmaßen nur der Punktestand als globales Leistungsmaß herangezogen.

Tabelle 20. Angenommene Richtung der Korrelationen für die in H3 erwarteten Zusammenhänge zwischen den Diagnosemaßen (Experiment 1)

		Pupillenweite	Fixationsdauer	EEG-Engagement	EEG-Frustration	**H3c** Punktestand
H3a	NASA-TLX-Gesamtscore	$\rho > 0$	$\rho < 0$	$\rho > 0$		$\rho < 0$
	NASA-TLX Frustration				$\rho > 0$	$\rho < 0$
	SAM Valenz				$\rho < 0$	$\rho > 0$
	SAM Arousal			$\rho > 0$	$\rho > 0$	$\rho < 0$
H3b	Punktestand	$\rho < 0$	$\rho > 0$	$\rho < 0$	$\rho < 0$	1

Grau: keine Hypothese definiert

4.2.8 Stichprobe und Versuchsdurchführung

An der experimentellen Untersuchung nahmen 14 Personen teil, wobei zwei Personen den Versuch vorzeitig abbrachen und in der Auswertung daher nicht berücksichtigt wurden. Bei den verbleibenden 12 Teilnehmern handelt es sich um 10 männliche und 2 weibliche Mitarbeiter des Fraunhofer FKIE im Alter zwischen 19 und 38 Jahren (Ø 30,1 Jahre). Die Versuchsteilnehmer hatten keine Vorerfahrungen mit Aufgaben der Luftraumüberwachung. Sie unterscheiden sich jedoch hinsichtlich ihrer Erfahrung mit computerbasierten Simulations- und Strategiespielen (vgl. Abschnitt 4.2.4 „Erfahrung“). Sechs Teilnehmer gaben sehr gute Kenntnisse an, die anderen sechs Teilnehmer gaben an, keine oder nur geringe Kenntnisse zu besitzen.

Die Durchführung der experimentellen Untersuchung fand in einem Laborraum des Fraunhofer FKIE statt und wurde von zwei wissenschaftlichen Mitarbeitern als Versuchsleiter mit Unterstützung einer Praktikantin durchgeführt. Da für die Untersuchung der Pupillenweite konstante Lichtbedingungen notwendig sind, wurde der Laborraum künstlich beleuchtet und die Fenster mit Vorhängen abgedunkelt. Abbildung 23a zeigt die räumliche Anordnung der Sitzplätze des Versuchsteilnehmers (vorne) und des Versuchsleiters (hinten). Der Sitzplatz des Versuchsleiters befand sich seitlich zu dem des Versuchsteilnehmers, so dass dieser bei Problemen oder Unklarheiten ansprechbar war.

Abbildung 23b zeigt den Arbeitsplatz des Versuchsteilnehmers von vorne. Die Experimentalaufgabe wurde an einem 24-Zoll-Monitor mit den Eingabegeräten Tastatur und Maus durchgeführt. Unterhalb des Monitors ist der verwendete Eyetracker Tobii X120 zu sehen. Die auditiven Stimuli (Alarme, Rechenaufgaben, Hintergrundlärm) wurden über In-Ear-Kopfhörer präsentiert (vgl. Abbildung 23b links). Das Emotiv EEG, das sich in Abbildung 23b rechts neben der Tastatur befindet, wurde vor Beginn der Untersuchung am Kopf des Probanden (siehe Abbildung 23a) angebracht.

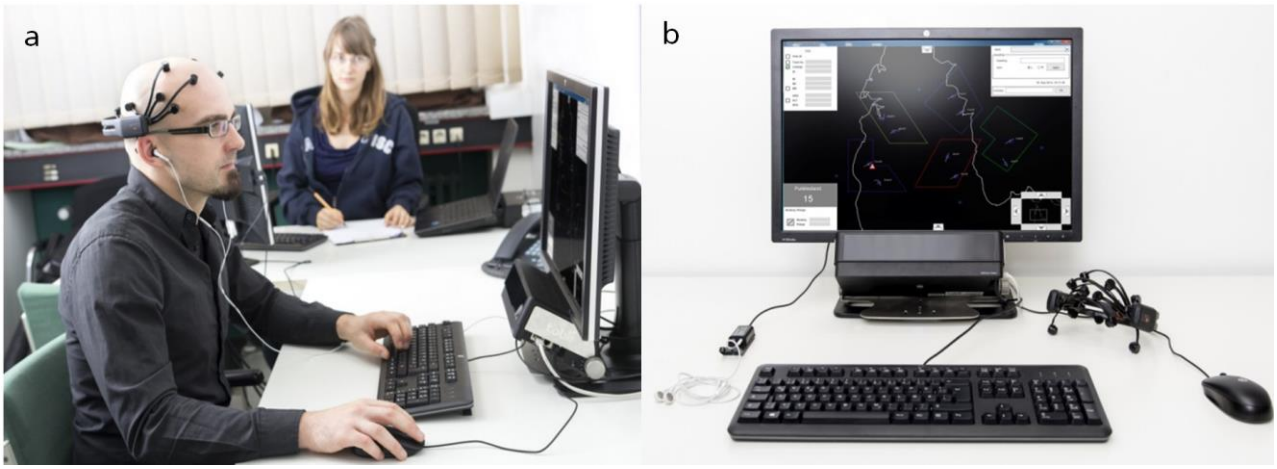


Abbildung 23. Versuchsaufbau (Experiment 1)

Jeder Versuchsteilnehmer nahm entsprechend dem Versuchsplan (siehe Abschnitt 4.2.2) an zwei Versuchssitzungen (A und B) teil. Die Versuchsdauer betrug pro Sitzung etwa 2 Stunden. Um Ermüdungseffekte gering zu halten und tageszeitabhängige Effekte zu kontrollieren, wurden die Sitzungen pro Teilnehmer an unterschiedlichen Tagen aber in etwa zur gleichen Uhrzeit durchgeführt. Der grundsätzliche Ablauf der Untersuchung war bei beiden Sitzungen identisch. Lediglich eine Einverständniserklärung (vgl. Anhang A.2) und ein Fragebogen zu Personenangaben (Alter, Geschlecht, Händigkeit, Erfahrung) wurden nur in der ersten Versuchssitzung zu Beginn ausgefüllt. In Tabelle 21 sind die wesentlichen Schritte der Versuchsabfolge pro Versuchssitzung zusammen mit der ungefähren Dauer aufgeführt.

Tabelle 21. Versuchsabfolge pro Versuchssitzung (Experiment 1)

Nr.	Aktion	Dauer
1	Einverständniserklärung und Fragebogen zur Person (nur im ersten Teilexperiment)	5 min
2	Anbringen/Kalibrieren der physiologischen Sensoren	20 – 30 min
3	Fragebogen zum Befinden und Vorabtests	15 – 20 min
4	Erhebung der Baseline	5 min
5	Instruktion und Übungsszenario	10 min
6	Erster Test mit Fragebögen NASA-TLX und SAM	15 min
7	Zweiter Test mit Fragebögen NASA-TLX und SAM	15 min
8	Dritter Test mit Fragebögen NASA-TLX und SAM	15 min
9	Vierter Test mit Fragebögen NASA-TLX und SAM	15 min
10	Abschließende mündliche Fragen	5 min

Das EEG-Headset wurde direkt zu Beginn der Untersuchung angebracht, da die EEG-Klassifikatoren laut Herstellerangaben einige Zeit benötigen, bis sie sich selbst kalibriert haben (vgl. Abschnitt 4.2.5 zum Emotiv EPOC^{TM} EEG). Als nächstes folgte die Kalibrierung des Eyetrackers sowie die Erfassung der individuellen Einflussfaktoren über Fragebögen und Leistungstests (vgl. Abschnitt 4.2.4). Anschließend wurde für die EEG-Metriken eine Baseline im

Ruhezustand aufgezeichnet[4]. Nach der Instruktion und der Bearbeitung eines Übungsszenarios folgten vier Testdurchläufe von je 10 Minuten Dauer, die jeweils unterschiedliche Versuchsbedingungen darstellten (vgl. Abschnitt 4.2.2). Nach jedem Test wurde der Versuchsteilnehmer gebeten, die Fragebögen NASA-TLX und SAM zu beantworten. Zum Abschluss der Untersuchung stellte der Versuchsleiter noch einige Fragen zur Untersuchung in mündlicher Form (vgl. Anhang A.8).

4.2.9 Vorgehen bei der Datenaufbereitung

Um die aufgezeichneten Daten inferenzstatistisch auswerten zu können, waren verschiedene Schritte der Datenaufbereitung notwendig, die im Folgenden näher erläutert werden.

Bereinigung

Die Datensätze der physiologischen Maße mussten zunächst von fehlerhaften Werten bereinigt werden. Bei den Emotiv-Metriken wird die Erfassung deaktiviert, wenn zu viel Rauschen vorhanden ist, um den Zustand korrekt zu berechnen. Jedoch werden in diesen Fällen in der Ausgabedatei konstante Werte ausgegeben. Da diese von den Analyseprogrammen als gültige Werte interpretiert werden, kann dies zu verzerrten Ergebnissen führen. Daher wurden Werte, die über mehrere Sekunden konstant blieben, aus den Datensätzen entfernt. Bei den Eyetracking-Daten entfiel eine Bereinigung, da keine Werte ausgegeben werden, wenn der Eyetracker nicht in der Lage ist, reliable Blickdaten aufzuzeichnen.

Es wurde außerdem geprüft, ob Tests einen so hohen Anteil fehlender Werte aufweisen, dass sie aus der Analyse ausgeschlossen werden müssen. Für die Eyetracking-Daten gibt Tobii Studio bei jeder abgeschlossenen Aufzeichnung einen Prozentwert für die Aufzeichnungsqualität an. Der Wert entspricht dem Anteil an Messvorgängen („samples“) pro Aufzeichnung, bei denen der Eyetracker valide Blickdaten erfassen konnte. Bei den EEG-Metriken ist kein Qualitätsindikator vorhanden. Der Anteil gültiger Werte pro Aufzeichnung musste daher manuell berechnet werden. Dem Vorgehen von Kardan & Cornati (2012) folgend wurde analysiert, welcher Grenzwert Tests mit besonders geringer Aufzeichnungsqualität ausschließt und dabei die Anzahl auszuschließender Tests möglichst gering hält. Auf Basis der in Anhang A.9 aufgeführten Diagramme zur Aufzeichnungsqualität wurde ein Anteil gültiger Werte von 30% als Grenzwert festgelegt. Für jede abhängige Variable gilt somit, dass Testdurchgänge mit einer Aufzeichnungsqualität unter 30% von der inferenzstatistischen Auswertung ausgeschlossen wurden. Dies führt bei der varianzanalytischen Auswertung (siehe Abschnitt 4.2.10) zu einer Reduzierung der Stichprobengröße, da nur Probanden mit vollständigen Datensätzen in die Analyse einbezogen werden können. Die Anzahl der pro Variable einbezogenen Versuchspersonen ist bei der Ergebnisdarstellung in Abschnitt 4.3.1, S.93 aufgeführt.

[4] Die Baseline wurde bei der Auswertung der Daten allerdings nicht herangezogen, da sich herausstellte, dass die aufgezeichneten Werte trotz des Ruhezustands noch starke Schwankungen aufwiesen. Es ist zu vermuten, dass die Selbstkalibrierung der Klassifikatoren zu diesem Zeitpunkt noch nicht abgeschlossen war.

Synchronisierung und Aggregierung

Da die Maße des Eyetrackers und des EEG sowie die Leistungsparameter (u.a. Auftreten und Dauer von Alarmen, Punktestand) über separate Schnittstellen und mit unterschiedlicher Frequenz (120 Hz beim Eyetracker, max. 43 Hz beim EEG) aufgezeichnet werden, war eine Synchronisierung und Aggregierung der Daten erforderlich.

Für die Synchronisierung wurde die in jeder Log-Datei ausgegebene Systemzeit herangezogen. Da in der Log-Datei des Experimentalsystems Beginn und Ende des Versuchsszenarios dokumentiert sind, wurden die Start- und Endzeiten von allen Datensätzen pro Szenario und pro Versuchsperson an die jeweilige Start- und Endzeit der Log-Datei des Experimentalsystems angeglichen. In einem weiteren Schritt wurden die Werte in allen Log-Dateien auf Sekundenbasis aggregiert. Da die Dauer der Szenarien pro Versuchsbedingung 10 Minuten betrug, resultierten für jede Variable 600 Werte pro Versuchsbedingung.

Standardisierung

Für die inferenzstatistische Auswertung war es erforderlich, die Werte jeder Variablen nach der beschriebenen Datenaufbereitung pro Person und Szenario zu einem Mittelwert zusammenzufassen. Bei den physiologischen und verhaltensbasierten Maßen ist jedoch zu beachten, dass diese an einem individuellen Referenzwert relativiert werden sollten, um interindividuelle Unterschiede auszugleichen (vgl. Anforderung 3 in Tabelle 7, S. 33). Hierfür wurde das Verfahren der z-Transformation angewendet. Für jede Person werden die Mittelwerte pro Versuchsbedingung X_i von dem Gesamtmittelwert $\overline{X}$ (Mittelwert aller Versuchsbedingungen) subtrahiert. Das Ergebnis wird durch die Standardabweichung s geteilt (vgl. Gleichung 2).

$$z = \frac{\bar{X} - X_i}{s}$$

Gleichung 2. Transformation der Mittelwerte in z-standardisierte Werte.

Eine z-Transformation bewirkt, dass die Verteilung einen Mittelwert von 0 und eine Standardabweichung von 1 erhält. Sie ermöglicht es somit, dass Maße mit unterschiedlicher Skalierung miteinander verglichen werden können (vgl. Wirtz & Nachtigall, 2002). Da die Transformation für jede Person einzeln vorgenommen wird, ist es allerdings nicht möglich, Effekte zwischen Personen, wie sie in H2b angenommen werden, zu untersuchen. Bei der Pupillenweite ist außerdem zu beachten, dass diese auch durch externe Faktoren (z.B. Lichtverhältnisse, Koffeinkonsum) beeinflusst wird, die sich bei den jeweiligen Versuchsteilnehmern zwischen dem ersten und dem zweiten Versuchstermin (trotz einer weitgehenden Abdunklung des Versuchsraums durch Gardinen) unterschieden haben könnten. Um einer Konfundierung entgegenzuwirken, wird für die Standardisierung der Pupillenweite daher nicht der Gesamtmittelwert sondern der Mittelwert pro Versuchstermin herangezogen. Eine Einschränkung besteht hierbei darin, dass die Auswirkungen der Lärmbedingung, die zwischen den Versuchsterminen variiert wurde, für die Pupillenweite nicht untersucht werden konnten.

4.2.10 Statistische Auswertung

Die statistische Auswertung erfolgte mit dem Statistikprogramm SPSS 20.0 (IBM Corp., 2011), die grafische Visualisierung der Ergebnisse mit Microsoft Excel 2010. Im Rahmen einer deskriptiven Analyse der Daten wurde zunächst untersucht, ob Auffälligkeiten (Extremwerte, fehlende Werte, ungleiche Varianzen, Abweichungen von der Normalverteilung) vorhanden sind, die bei der inferenzstatistischen Auswertung zu berücksichtigen sind.

Grundsätzlich wird für alle Verfahren ein Signifikanzniveau von $\alpha = .05$ zugrunde gelegt, wobei zur Vermeidung einer Alpha-Fehler-Kumulierung bei Testung einer Hypothese durch mehrere Variablen der gleichen Methodenkategorie die Bonferroni-Holm-Korrektur (Holm, 1979) angewendet wurde. Neben der Signifikanz wurde bei der Analyse und Bewertung der Ergebnisse auch die Effektstärke als Maß für die praktische Relevanz berücksichtigt. Bei korrelativen Zusammenhängen wurde die Klassifizierung der Korrelationswerte *r* in schwache, mittlere und starke Effekte nach Cohen (1988) herangezogen:

$$|r| \approx 0{,}1\text{: schwacher Effekt; } |r| \approx 0{,}3\text{: mittlerer Effekt } |r| \approx 0{,}5\text{: starker Effekt}$$

<u>Auswertungen zu H1</u>

Zur Testung der statistischen Hypothesen zu Hypothese 1 wurde das Verfahren der Varianzanalyse mit Messwiederholung angewendet. Die drei experimentell variierten Faktoren *Area, Kooperativität* und *Lärm* gehen dabei als Messwiederholungsfaktoren mit je zwei Faktorstufen in die Analyse ein. Als abhängige Variablen (AV) fungieren die Indikatoren des Nutzerzustands (vgl. Abschnitt 4.2.5 und 4.2.6). Die Auswertung erfolgte über univariate Varianzanalysen (ANOVA). Auf eine multivariate Auswertung der AV im Rahmen einer MANOVA (Multivariate Analysis of Variance) wurde verzichtet, da die Anzahl der im Modell enthaltenen AV die Teststärke in Anbetracht der Stichprobengröße zu sehr verringern würde (vgl. Empfehlungen zur Stichprobengröße in Tabachnick & Fidell, 2007, S. 250).

Die univariate Varianzanalyse (ANOVA) kann nach Bortz (2005, S.287) bei gleich großen Gruppen, wie sie im Messwiederholungsdesign gegeben sind, als robust gegenüber Verletzungen der Voraussetzungen (z.B. der Normalverteilungannahme) angesehen werden. Dennoch wurden bei signifikanten Abweichungen von der Normalverteilung im Kolmogorov-Smirnov-Test auch nonparametrische Verfahren (z.B. Wilcoxon-Signed-Rank-Test, Wilcoxon, 1945) zur statistischen Absicherung einzelner Effekte herangezogen. Für die ANOVA mit Messwiederholung gilt außerdem die Voraussetzung der Varianzhomogenität bzw. Homogenität der Korrelationen zwischen Faktorstufen. Bei lediglich zwei Stufen pro Faktor, wie sie in der Untersuchung vorliegen, ist diese Forderung jedoch bedeutungslos, da jeweils nur eine Korrelation pro Faktor berechnet werden kann (Bortz, 2005, S. 354).

<u>Auswertungen zu H2</u>

Bei Hypothese 2 soll der Einfluss jedes individuellen Faktors auf den subjektiven Nutzerzustand und die Leistung statistisch untersucht werden (vgl. Abschnitt 4.2.7). In diesem Fall kann die Hypothesentestung sowohl über bivariate Regressionen als auch korrelativ erfolgen, da der Steigungskoeffizient β mit der Produkt-Moment-Korrelation *r* identisch ist (Bortz, 2005, S. 207).

Voraussetzung der Regressionsanalyse sowie der Produkt-Moment-Korrelation ist jedoch u.a. das Vorliegen metrischer unabhängiger und abhängiger Variablen. Da für die individuellen Faktoren kein metrisches Skalenniveau vorausgesetzt werden kann, wurde die Spearman-Rho Korrelation als nonparametrisches Verfahren zur Berechnung der Zusammenhänge herangezogen.

Zu beachten ist, dass Kausalaussagen zwar durch Nullkorrelationen falsifiziert werden können, dass durch das Vorliegen statistisch signifikanter Assoziationen jedoch noch keine Kausalität im Sinne einer Ursache-Wirkung-Relation nachgewiesen ist (Bortz & Döring, 2006, S. 518). Wichtig ist überdies zu prüfen, ob eine gegenläufige Verursachungsrichtung in Frage kommt, oder ob der Zusammenhang auf einer Scheinkorrelation beruhen könnte. Inwiefern diese Aspekte in der vorliegenden Untersuchung in Betracht kommen, wird in Abschnitt 4.4.2 diskutiert.

Auswertungen zu H3

Hypothese 3 wurde ebenfalls korrelativ untersucht. Da es sich bei den Indikatoren um metrische Variablen handelt, wurden Produkt-Moment-Korrelationen berechnet. Eine Korrektur von Ausreißerwerten war nicht erforderlich, jedoch reduzierte sich die Fallzahl bei einigen physiologischen Parametern durch den Ausschluss fehlerhafter Werte (vgl. Abschnitt 4.2.9). Die Auswertungen erfolgten sowohl auf Gruppenebene unter Einbezug der Daten von allen Teilnehmern in allen Bedingungen als auch für jede Person einzeln (individuelle Ebene). In die Individualanalyse wurden lediglich die personenspezifischen Werte der acht Versuchsbedingungen einbezogen. Bei den Korrelationen auf Gruppenebene ist zu beachten, dass mehrere Messwerte pro Person in die Analysen eingehen, so dass Unterschiede zwischen und innerhalb von Personen konfundiert sind (vgl. Bland & Altman, 1995). Durch die z-Standardisierung der Eyetracking- und EEG-Maße wurden Zwischensubjekteffekte für diese Variablen jedoch bereits neutralisiert. Dies gilt allerdings nicht für die subjektiven Maße und die Leistungsmaße. Bland & Altman (1995) schlagen bei Vorliegen von mehreren Messwerten pro Person vor, die Versuchspersonen in Form von Dummy-Variablen in die Analyse einzubeziehen. Effekte, die auf Unterschiede zwischen Personen zurückgehen, können so aus dem Zusammenhang auspartialisiert werden. Bei den Ergebnissen auf Gruppenebene wird daher neben der Gesamtkorrelation auch die Partialkorrelation aufgeführt, bei der Zwischensubjekteffekte durch die Hinzunahme der Versuchspersonen als Dummy-Variablen auspartialisiert wurden.

4.3 Ergebnisse der experimentellen Untersuchung

In den folgenden Abschnitten werden die Ergebnisse der inferenzstatistischen Auswertung für jede der drei Forschungshypothesen und ihre Subhypothesen berichtet.

4.3.1 Einfluss der Anforderungssituation auf den diagnostizierten Nutzerzustand (Hypothese 1)

Hypothese 1 nimmt an, dass die drei experimentell variierten Merkmale der Anforderungssituation (Anzahl Areas, Kooperativität und Lärm) den diagnostizierten Nutzerzustand beeinflussen. Entsprechend der drei Subhypothesen wird diese Annahme gesondert für die subjektiven Maße (H1a), die physiologischen und verhaltensbasierten Maße (H1b) und die Leistungsmaße (H1c) untersucht.

H1a: Subjektive Maße

In Tabelle 22 sind die Ergebnisse der varianzanalytischen Auswertung für den *NASA-TLX-Gesamtscore* (arithmetisches Mittel der sechs Subskalen des NASA-TLX), die Subskala *Frustration* und die SAM-Dimensionen *Valenz und Arousal* dargestellt. Neben der Signifikanz ist auch das partielle Eta² (η_p^2) als Effektstärkemaß angegeben. Mittelwerte und Standardfehler pro Variable und Faktorstufe sind in Abbildung 24 in einem Balkendiagramm veranschaulicht.

Tabelle 22. Ergebnisse der varianzanalytischen Auswertung bezüglich der Fragebögen NASA-TLX und SAM (Experiment 1)

	Haupteffekt Areas			**Haupteffekt Kooperativität**			**Haupteffekt Lärm**		
	$F(1,11)$	p	η_p^2	$F(1,11)$	p	η_p^2	$F(1,11)$	p	η_p^2
NASA-TLX Gesamtscore	51.80	<.001	.83	20.26	.001	.65	.66	.44	.06
NASA-TLX Frustration	24.94	<.001	.69	11.56	<.01	.51	.17	.69	.02
SAM Valenz	7.05	<.05	.39	3.31	.1	.23	1.98	.19	.15
SAM Arousal	16.5	<.01	.60	.31	.59	.03	.11	.74	.01

Anmerkung: alle *p*-Werte <.05 sind bei Anwendung der Bonferroni-Holm-Korrektur signifikant.

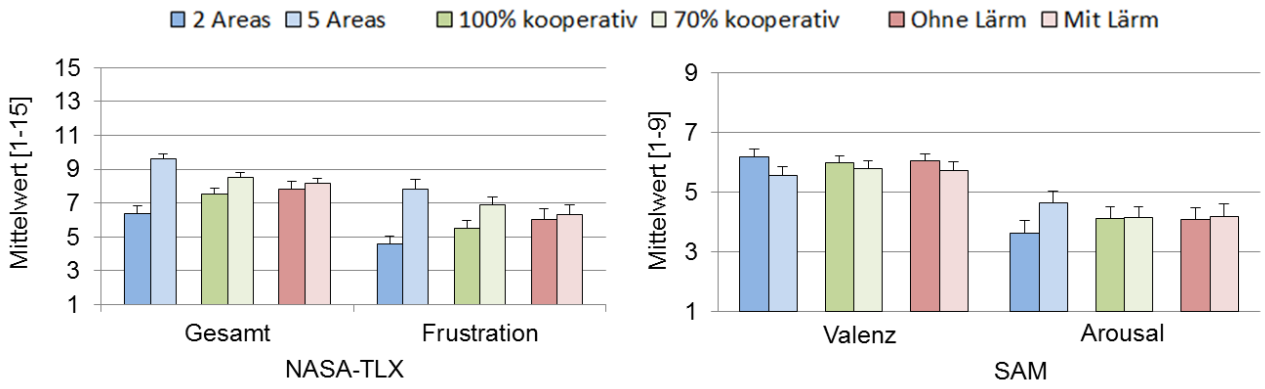


Abbildung 24. Mittelwerte pro Faktorstufe für die subjektiven Bewertungen in Experiment 1 (Fehlerbalken: Standardfehler)

Analyse der Ergebnisse zum NASA-TLX

Aus Tabelle 22 wird ersichtlich, dass der *NASA-TLX-Gesamtscore* signifikante Haupteffekte in Bezug auf die Faktoren Anzahl Areas und Kooperativität aufweist. Der Haupteffekt für den Faktor Lärm erweist sich hingegen als nicht signifikant. Wie aus Abbildung 24 hervorgeht, liegt der Gesamtscore bei fünf zu überwachenden Areas höher als bei zwei Areas. Dies entspricht somit der in H1a formulierten Annahme, dass eine größere Anzahl an Areas die Belastung erhöht und sich in einer höheren Beanspruchung auswirkt. Außerdem zeigt sich, dass die Beanspruchung bei geringer Kooperativität höher bewertet wird als bei hoher Kooperativität der Kontakte. Die Effektstärke fällt für den Faktor *Areas* mit η_p^2 =.83 sehr hoch aus und liegt für den Faktor Kooperativität mit η_p^2 = .65 etwas niedriger.

Die Subskala *Frustration* des NASA-TLX weist ebenfalls signifikante Haupteffekte für die Faktoren Anzahl Areas und Kooperativität auf. Wie für den Gesamtscore zeigt sich auch hier, dass die Frustration bei 5 Areas höher bewertet wird als bei 2 Areas und dass sie bei geringer

Kooperativität höher bewertet wird als bei hoher Kooperativität. Dies bestätigt die Annahme, dass eine verminderte Kooperativität der Kontakte den Grad der Frustration erhöht. Aus den η_p^2-Werten in Tabelle 22 geht aber auch hervor, dass der Effekt für den Faktor Anzahl Areas größer ausfällt als für Kooperativität.

Bei der Interpretation von Haupteffekten müssen auch die Interaktionen zwischen den Faktoren geprüft werden. Signifikante Haupteffekte können dann global, d.h. unabhängig von einem anderen Faktor interpretiert werden, wenn keine Interaktion oder eine ordinale Interaktion vorliegt. Bei einer hybriden Interaktion gilt dies nur für einen Faktor (Bort & Döring, 2006, S. 534). Für den *NASA-TLX-Gesamtscore* stellte sich einzig die Interaktion zwischen Kooperativität und Lärm als signifikant heraus ($F(1,11)=5.45$; $p<.05$; $\eta_p^2 =.33$). Eine Detailanalyse zeigt, dass eine hybride Interaktion vorliegt (vgl. Interaktionsdiagramme in Anhang A.11), wobei sich die Interaktion nicht auf die Interpretation des Haupteffekts für Kooperativität auswirkt, da die Beanspruchung sowohl mit als auch ohne Lärm bei geringer Kooperativität im Durchschnitt höher bewertet wurde als bei hoher Kooperativität. In Bezug auf die Subskala *Frustration* ergaben sich keine signifikanten Interaktionen zwischen den Faktoren. Die signifikanten Haupteffekte für Anzahl Areas und Kooperativität können auch hier global interpretiert werden. Ebenso erwiesen sich die Interaktionen 2. Ordnung für den NASA-TLX-Gesamtscore ($F(1,11)=0.25$; $p=.63$) und die Subskala Frustration ($F(1,11)=0.38$; $p=.55$) als nicht signifikant.

Analyse der Ergebnisse zum SAM

Bei den Dimensionen des SAM treten signifikante Haupteffekte für den Faktor Anzahl Areas sowohl hinsichtlich Valenz als auch Arousal auf (vgl. Tabelle 22). Entsprechend der in Abschnitt 4.2.7 getroffenen Annahme wird das Arousal bei 5 Areas höher bewertet als bei 2 Areas (vgl. Abbildung 24). Zudem wird – konsistent mit den Ergebnissen zur Frustration – die Valenz bei 5 Areas negativer bewertet als bei 2 Areas. Die Haupteffekte für Kooperativität und Lärm ebenso wie die Interaktionen sind jedoch bei beiden SAM-Diensionen nicht signifikant. Somit lässt sich die in H1a getroffene Annahme, dass sich Unterschiede im emotionalen Zustand insbesondere zwischen geringer und hoher Kooperativität zeigen, durch den SAM nicht bestätigen.

H1b: Physiologische und verhaltensbasierte Maße

Für die Untersuchung von Hypothese H1b wurden Varianzanalysen für die Eyetracking-Maße *Pupillenweite* und *Fixationsdauer* sowie für die Klassifikatoren *EEG-Engagement* und *EEG-Frustration* berechnet (vgl. Tabelle 23). Bei EEG-Engagement mussten fünf Versuchspersonen aufgrund fehlender Werte ausgeschlossen werden, bei EEG-Frustration eine Person und bei den Eyetracking-Maßen zwei Personen. Da es sich um mehrere unabhängige Varianzanalysen handelt, wurde in Hinblick auf eine Kumulierung des α-Fehlers zusätzlich geprüft, ob die Ergebnisse nach Anwendung der Bonferroni-Holm-Korrektur signifikant ausfallen. In Abbildung 25 sind für jede Variable die Mittelwerte und Standardfehler pro Faktorstufe als Balkendiagramme aufgeführt.

Für die *Fixationsdauer* kann ein signifikanter Haupteffekt für den Faktor Anzahl Areas festgestellt werden. Wie in Hypothese H1b angenommen, geht die Bedingung mit 5 Areas mit einer kürzeren Fixationsdauer einher als die Bedingung mit 2 Areas. Die *Pupillenweite* weist für den Faktor Anzahl Areas einen Effekt mittlerer Größe auf, der jedoch nach Anwendung der Bonferroni-Holm-

Korrektur nicht mehr signifikant ausfällt. Die Ergebnisse entsprechen allerdings den Erwartungen, dass die Pupillenweite bei 5 Areas im Durchschnitt größer ist als bei 2 Areas (vgl. Abbildung 25).

Für *EEG-Engagement* zeigt sich ein signifikanter Haupteffekt in Bezug auf den Faktor Anzahl Areas. Wie in Abbildung 25 zu sehen ist, liegt das Engagement in der Bedingung mit 5 Areas signifikant höher als in der Bedingung mit 2 Areas. Dies ist somit konform zu der in H1b getroffenen Annnahme einer höheren Beanspruchung bei 5 Areas im Vergleich zu 2 Areas. Das Engagement fällt zudem in der Bedingung mit Lärm erwartungsgemäß etwas höher aus als in der Bedingung ohne Lärm. Allerdings erweist sich dieser Effekt in der Varianzanalyse als nicht signifikant. Des Weiteren wurden Unterschiede in der Frustration insbesondere in Bezug auf den Faktor Kooperativität erwartet. Diese Annahme kann durch den Klassifikator *EEG-Frustration* nicht bestätigt werden. Hingegen zeigt sich in Bezug auf den Faktor Lärm, dass die Bedingung ohne Lärm mit einer höheren Frustration einhergeht als die Bedingung mit Lärm. Dieser Unterschied erweist sich nach Bonferroni-Holm-Korrektur zwar nicht mehr als signifikant, ist aber dennoch unerwartet. Sowohl bei den EEG- als auch bei den Eyetracking-Maßen konnten keine signifikanten Interaktionen zwischen Faktoren festgestellt werden, so dass die Effekte global interpretiert werden können.

Tabelle 23. Ergebnisse der varianzanalytischen Auswertung bezüglich der Eyetracking- und EEG-Maße in Experiment 1

		Haupteffekt Areas			**Haupteffekt Kooperativität**			**Haupteffekt Lärm**		
	df	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Pupillenweite	1,9	6.38	<.05[a]	.42	2.21	.17	.20		-	
Fixationsdauer	1,9	24.75	<.01	.73	2.31	.16	.20	.03	.87	.00
Engagement	1,6	10.64	<.05	.64	.32	.59	.05	1.87	.22	.24
Frustration	1,10	4.12	.07	.29	.39	.55	.04	6.57	<.05 [a]	.40

Anmerkung: [a] bei Anwendung der Bonferroni-Holm-Korrektur nicht signifikant.

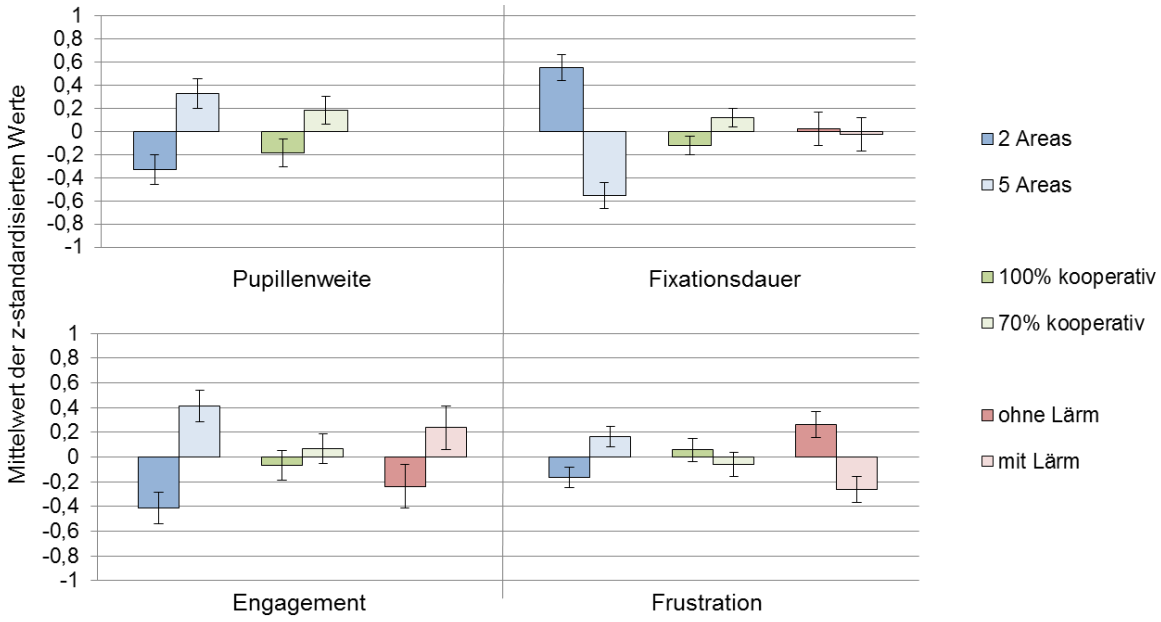


Abbildung 25. Mittelwerte pro Faktorstufe für die Eyetracking- und EEG-Maße in Experiment 1 (Fehlerbalken: Standardfehler)

H1c: Leistungsmaße

Um Hypothese H1c zu prüfen wurden Varianzanalysen für die Leistungsmaße *Dauer der Alarme*, die *Prozentzahl ausgelassener (Rechen-)Aufgaben* und den *Punktestand* berechnet. Eine hohe Leistung wird dabei angezeigt durch eine geringe Dauer der Alarme, einer niedrige Prozentzahl ausgelassener Aufgaben und einen hohen Punktestand.

In Tabelle 24 sind die Ergebnisse der varianzanalytischen Auswertung für diese Maße zusammengefasst. Da die Verteilungen für die Variablen Dauer Alarme und Ausgelassene Aufgaben linkssteil sind, wurde die Signifikanz für diese Variablen zusätzlich mit dem Wilcoxon-Signed-Rank-Test untersucht. Abbildung 26 veranschaulicht die jeweiligen Unterschiede zwischen den Faktorstufen als Balkendiagramm.

Tabelle 24. Ergebnisse der varianzanalytischen Auswertung bezüglich der Leistungsmaße (Experiment 1)

	Haupteffekt Areas			**Haupteffekt Kooperativität**			**Haupteffekt Lärm**		
	$F(1,11)$	p	η_p^2	$F(1,11)$	p	η_p^2	$F(1,11)$	p	η_p^2
Dauer Alarme	21.51	.001	.66	15.39	<.01	.58	3.08	.11	.22
Wilcoxon	z = -5.02	<.001		z = -4.21	<.001		z = -2.3	<.001	
Ausg. Aufgaben	33.23	<.001	.75	9.17	<.05	.46	3.08	.11	.22
Wilcoxon	z = -5.05	<.001		z = -2.26	<.05		z = -2.12	<.05	
Punktestand	102.72	<.001	.90	25.08	<.001	.70	0.58	.46	.05

Alle p-Werte <.05 sind bei Anwendung der Bonferroni-Holm-Korrektur signifikant.

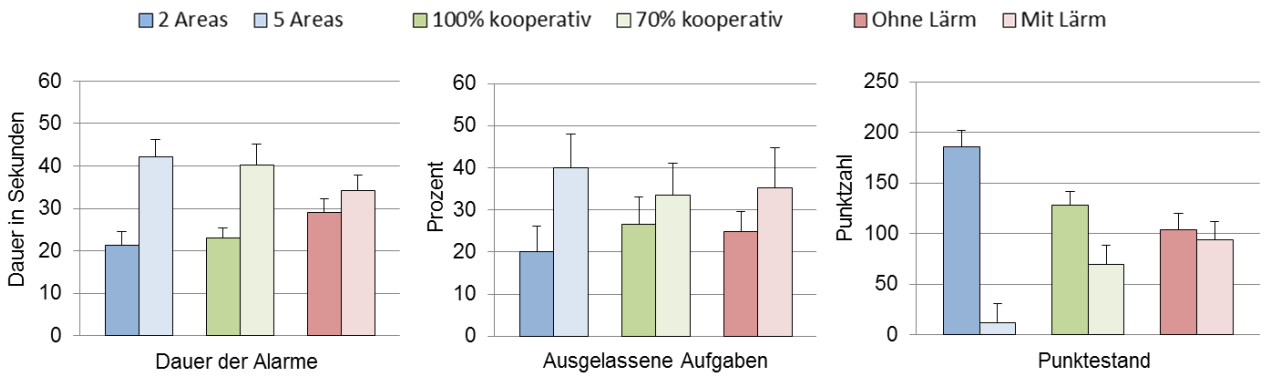


Abbildung 26. Mittelwerte pro Faktorstufe für die Leistungsmaße in Experiment 1 (Fehlerbalken: Standardfehler)

Es zeigt sich, dass alle Leistungsmaße signifikante Haupteffekte in Bezug auf die Faktoren Areas und Kooperativität aufweisen. Bei den Interaktionen ergab sich nur für die Variable Ausgelassene Rechenaufgaben eine signifikante ordinale Interaktion zwischen Anzahl Areas und Kooperativität ($F(1,11)=4.96$; $p<.05$; $\eta_p^2=.31$; siehe Anhang A.11, Abbildung 54) Bei Berechnung des Wilcoxon-Signed-Rank-Tests für die Variablen Dauer der Alarme und Ausgelassene Rechenaufgaben erwies sich – im Unterschied zur Varianzanalyse – auch der Unterschied für Lärm als signifikant.

Wie in Hypothese H1c angenommen, bestätigen die Ergebnisse, dass bei 2 Areas und bei hoher Kooperativität eine bessere Leistung erzielt wurde als bei 5 Areas und bei geringer Kooperativität. Diese äußert sich in einer kürzeren Dauer von Alarmen, weniger ausgelassenen Rechenaufgaben als auch in einem höheren Punktestand (vgl. Abbildung 26).

Für Lärm fallen die Unterschiede geringer aus, aber auch hier wird ohne Lärm eine bessere Leistung erzielt als mit Lärm, was den Annahmen in H1c entspricht.

4.3.2 Einfluss individueller Faktoren auf den Nutzerzustand und auf die Leistung (Hypothese 2)

In Hypothese 2 wird angenommen, dass individuelle Faktoren zu interindividuellen Unterschieden im Nutzerzustand und der Leistung beitragen. Konkret wurde dies untersucht für die *Fähigkeiten* (Leistung im Rechentest und Links-Rechts-Test), die *Erfahrung mit Computerspielen*, die *Müdigkeit* und die *Motivation*. In Tabelle 25 sind die Spearman-Rho-Korrelationen der individuellen Faktoren mit dem NASA-TLX-Gesamtscore, dem Arousal (SAM) und den Leistungsmaßen aufgeführt. Da es sich um gerichtete Hypothesen handelt, erfolgte die Signifikanztestung einseitig.

Tabelle 25. Spearman-Rho-Korrelationskoeffizienten für die Korrelationen zwischen individuellen Faktoren und Indikatoren des Nutzerzustands (Experiment 1)

	Diagnosemaß	L-R-Test	Rechentest	Erfahrung	Müdigkeit	Motivation
H2a	NASA-TLX	**-.21**	-.02	**-.31**		**.16**
	Arousal (SAM)				**-.08**	**.30**
H2c	Dauer Alarme	**-.03**	**-.47***	**-.57****	**.11**	.14
	Ausg. Rechenaufgaben	**-.27**	**-.52****	**-.34**	**.31**	**-.51****
	Punktestand	**.48****	**.63****	**.59****	**-.33**	**.43***

** p<.01; * p<.05 bei einseitigem Testen; fettgedruckt: in erwarteter Richtung, Fallzahl: N=24 (1. Durchgang & 2. Durchgang)

In Bezug auf die über den NASA-TLX erfasste subjektive Beanspruchung und das Arousal (SAM) erweist sich keine Korrelation als signifikant. Allerdings fallen die Korrelationen zwischen dem NASA-TLX und dem Links-Rechts-Test, der Erfahrung und der Motivation in erwarteter Richtung aus und weisen auf zumindest schwache bis mittlere Zusammenhänge hin. Für das *Arousal* zeigt sich ein erwartungskonform positiver Zusammenhang mit der Motivation und ein sehr geringer negativer Zusammenhang mit der Müdigkeit. Im Gegensatz zu den subjektiven Maßen ergeben sich für die Leistungsmaße deutlich höhere und zumeist auch signifikante Korrelationen in erwarteter Richtung mit den individuellen Faktoren. Insbesondere die Korrelationen für den Rechentest und die Erfahrung weisen auf starke Zusammenhänge mit der Leistung hin.

4.3.3 Zusammenhänge zwischen Diagnosemaßen des Nutzerzustands (Hypothese 3)

In Hypothese 3 wird angenommen, dass Diagnosemaße, die das gleiche Konstrukt erfassen, korrelieren sollten. Im Folgenden werden die Ergebnisse für die drei Subhypothesen auf Gruppen- und auf Individualebene analysiert.

<u>H3a: Zusammenhänge zwischen subjektiven und physiologischen Maßen</u>

In Tabelle 26 sind die Produkt-Moment-Korrelationen zwischen den subjektiven Maßen und den Eyetracking-und EEG-Maßen auf Gruppenebene dargestellt, für die in H3a statistische Hypothesen formuliert wurden. Neben den Korrelationen r sind auch die Partialkorrelationen aufgeführt (r_{part}),

bei denen der Einfluss von Zwischensubjekteffekten herauspartialisiert wurde. Es ist ersichtlich, dass die Maße Pupillenweite und EEG-Engagement positiv und die Fixationsdauer negativ mit dem NASA-TLX-Gesamtscore korrelieren, was den Annahmen in Abschnitt 4.2.7 entspricht. Im Vergleich fallen die Partialkorrelationen geringfügig höher aus als die unkorrigierten Korrelationen. Sie weisen für diese drei Maße auf starke signifikante Zusammenhänge mit der subjektiven Beanspruchung hin. Anders verhält es sich bei den Zusammenhängen zwischen EEG-Frustration und NASA-TLX-Frustration, Valenz (SAM) und Arousal (SAM). Hier sind keine oder nur sehr schwache Zusammenhänge zu verzeichnen. Für Arousal (SAM) ist dieser sogar entgegen den Erwartungen schwach negativ. Ein geringer aber zumindest positiver Zusammenhang liegt für EEG-Engagement und Arousal (SAM) vor.

Tabelle 26. Produkt-Moment-Korrelationen zwischen subjektiven Maßen und den Eyetracking-/EEG-Maßen auf Gruppenebene unkorrigiert (r) und nach Bereinigung von Zwischensubjekteffekten (r_{part}) – Experiment 1

	Pupillenweite		**Fixationsdauer**		**EEG-Engagement**		**EEG-Frustration**	
	r	r_{part}	*r*	r_{part}	*r*	r_{part}	*r*	r_{part}
NASA-TLX Gesamt Fallzahl	.42** 80	.45** 80	-.41** 80	-.45** 80	.48** 70	.51** 70	-	-
NASA-TLX Frustration Fallzahl	-	-	-	-	-	-	.09 88	.1 88
SAM Valenz Fallzahl	-	-	-	-	-	-	.01 88	.01 88
SAM Arousal Fallzahl	-	-	-	-	.16 70	.19 70	-.12 88	-.16 88

** $p<.01$; * $p<.05$ bei einseitigem Testen.

In Abbildung 27 sind die für jeden Teilnehmer ermittelten Individualkorrelationen zwischen dem NASA-TLX-Gesamtscore und dem EEG-Engagement, der Pupillenweite und der Fixationsdauer sowie zwischen der Subskala Frustration und EEG-Frustration als Balkendiagramme dargestellt. Personen mit fehlenden Werten in mehr als einer Versuchsbedingung wurden aus den Analysen auf Gruppen- und auf Individualebene ausgeschlossen.

Bei den Individualkorrelationen in Abbildung 27 zeigt sich, dass der NASA-TLX-Gesamtscore bei allen Teilnehmern positiv mit der *Pupillenweite* (vgl. Abbildung 27a) und dem *EEG-Engagement* (vgl. Abbildung 27c) korreliert. Zu beachten ist, dass sich aufgrund der geringen Fallzahl (n=8) nur sehr hohe Individualkorrelationen als signifikant erweisen. Bei EEG-Engagement sind es vier Personen und bei der Pupillenweite drei Personen, die signifikante Korrelationen mit dem NASA-TLX-Gesamtscore aufweisen. Mit r >.7 liegen sie deutlich über dem Niveau auf Gruppenebene. Einige wenige Individualkorrelationen fallen hingegen vergleichsweise niedrig aus und weisen nur auf schwache Zusammenhänge hin (vgl. VP 4 und VP 10 bei der Pupillenweite und VP 8 bei EEG-Engagement).

Für die *Fixationsdauer* können auf Individualebene größtenteils negative Korrelationen mit dem NASA-TLX-Gesamtscore festgestellt werden (vgl. Abbildung 27b), wobei diese insgesamt etwas schwächer sind als bei der Pupillenweite und EEG-Engagement. Nur eine Individualkorrelation erweist sich als signifikant. Entgegen den Erwartungen korreliert die Fixationsdauer bei zwei Teilnehmern positiv mit dem NASA-TLX-Gesamtscore. Somit zeigen sich auch hier deutliche interindividuelle Unterschiede in Höhe und Richtung der Korrelationen.

Für *EEG-Frustration* sind in Abbildung 27d exemplarisch die Individualkorrelationen mit der Subskala Frustration des NASA-TLX aufgeführt. Wie sich zeigt, fallen diese sehr heterogen aus. Sie reichen von hoch positiv über schwach negativ bis hin zu hoch negativ. Somit kann auch bei den Individualkorrelationen, ähnlich wie auf Gruppenebene, kein eindeutiger Zusammenhang zwischen EEG-Frustration und der subjektiven Bewertung festgestellt werden.

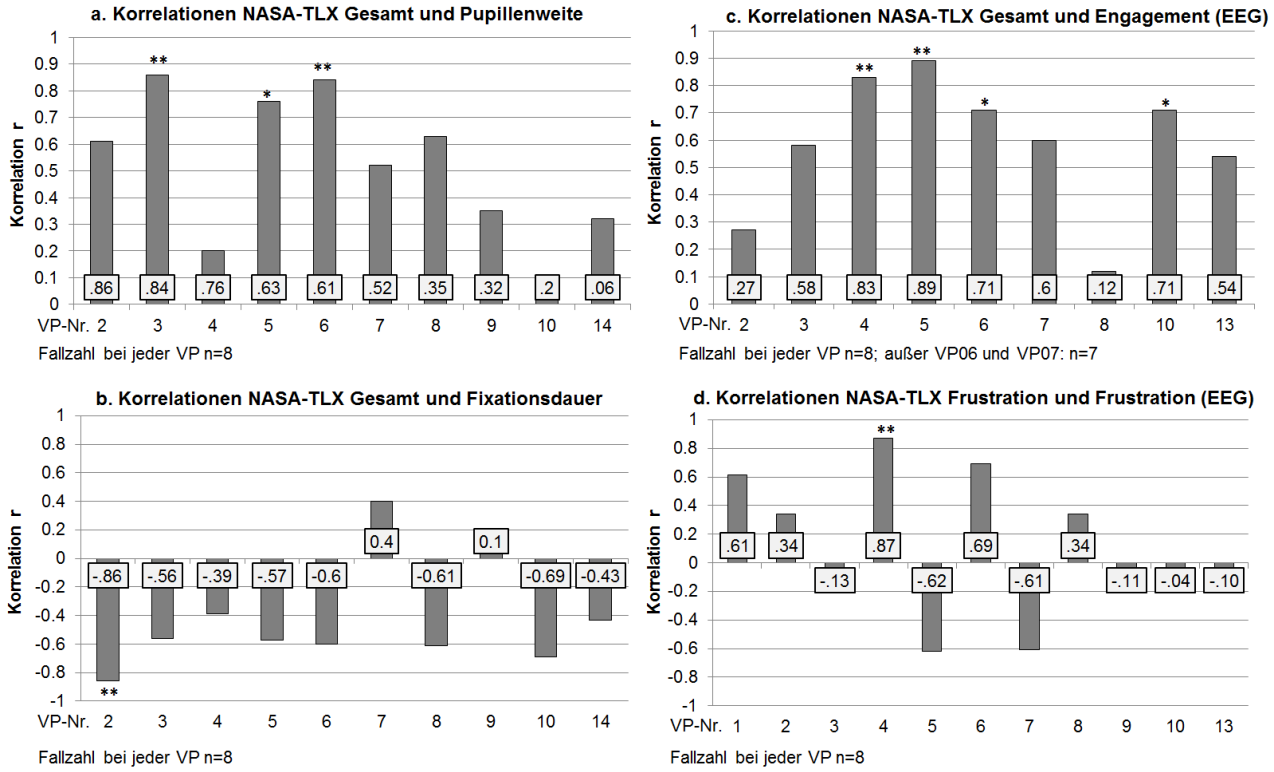


Abbildung 27. Individualkorrelationen pro Versuchsperson (VP) zwischen dem NASA-TLX-Gesamtscore und der Pupillenweite (a), der Fixationsdauer (b) und dem EEG-Engagement (c) sowie zwischen der Subskala Frustration und EEG-Frustration (d) (p<.01; * p<.05 bei einseitigem Testen) – Experiment 1**

Insgesamt kann konstatiert werden, dass Hypothese H3a für die *Pupillenweite*, die *Fixationsdauer* und das *EEG-Engagement*, bestätigt werden kann, da diese Maße auf Gruppenebene signifikant und in erwarteter Richtung mit der subjektiven Beanspruchung korrelieren und sich dies auch bei den meisten Individualkorrelationen widerspiegelt. Dabei ist zu beachten, dass die Individual-korrelationen zumeist zwar höher ausfallen als auf Gruppenebene, dass sich die Höhe der Zusammenhänge jedoch zwischen den Personen teilweise stark unterscheidet. Für *EEG-Frustration* kann die Hypothese, dass der Klassifikator mit subjektiven Maßen zur Frustration in erwarteter Weise korreliert, sowohl auf Gruppen- als auch auf Individualebene nicht bestätigt werden.

H3b: Zusammenhänge zwischen Leistungsmaßen und physiologischen/verhaltensbasierten Maßen

In Bezug auf H3b wurden die Zusammenhänge der betrachteten Eyetracking- und EEG-Maße mit der Leistung untersucht, wobei der *Punktestand* als Leistungsmaß herangezogen wurde. Tabelle 27 führt die Produkt-Moment-Korrelationen mit dem Punktestand auf Gruppenebene auf.

In den Spalten r_{part} sind die Partialkorrelationen aufgeführt, aus der der Einfluss von Zwischen-subjekteffekten auspartialisiert wurde. Die Korrelationen fallen erwartungskonform für die Pupillenweite, EEG-Engagement und EEG-Frustration negativ und für die Fixationsdauer positiv

aus. Die Partialkorrelationen weisen im Vergleich durchweg höhere Zusammenhänge auf als die nicht bereinigten Korrelationen. Mit Ausnahme von EEG-Frustration erweisen sie sich als signifikant (p<.01).

Tabelle 27. Produkt-Moment-Korrelationen zwischen dem Punktestand und den Eyetracking-/EEG-Maßen auf Gruppenebene unkorrigiert (r) und nach Bereinigung von Zwischensubjekteffekten (r_{part}) – Experiment 1

	Pupillenweite		Fixationsdauer		EEG-Engagement		EEG-Frustration	
	r	r_{part}	r	r_{part}	r	r_{part}	r	r_{part}
Punktestand	-.36**	-.40**	..45**	.50**	-.40**	-.44**	-.12	-.14
Fallzahl	80	80	80	80	70	70	88	88

** p<.01; * p<.05 bei einseitigem Testen.

Zusammenhänge auf Individualebene

Abbildung 28 stellt die Individualkorrelationen zwischen den Eyetracking- und EEG-Maßen mit dem Punktestand für jede Versuchsperson als Balkendiagramm dar.

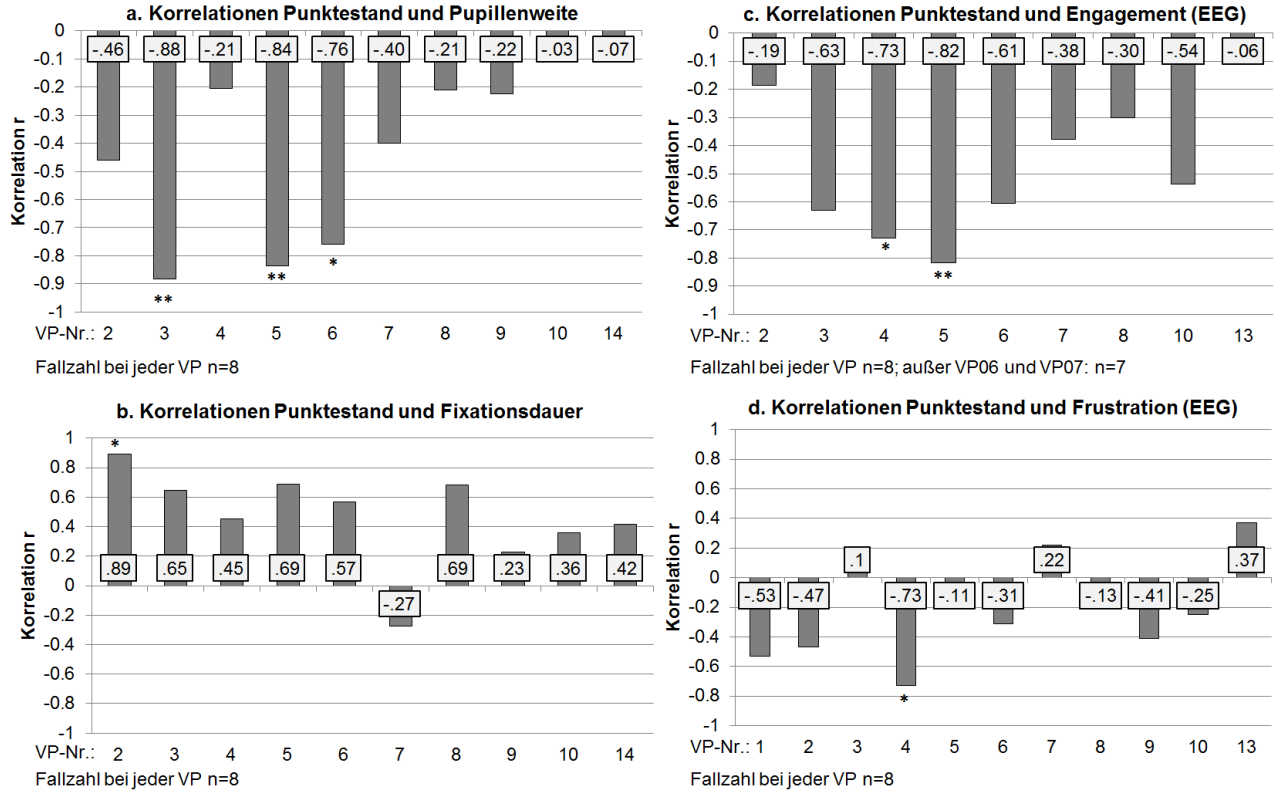


Abbildung 28. Individualkorrelationen pro Versuchsperson (VP) zwischen den Eyetracking- und EEG-Maßen und dem Punktestand (p<.01; * p<.05 bei einseitigem Testen) – Experiment 1**

Für die Pupillenweite und das EEG-Engagement können erwartungsgemäß durchweg negative Korrelationen mit dem Punktestand festgestellt werden (vgl. Abbildung 28a und c). Bei einigen Probanden weisen diese mit r > .5 auf einen starken Effekt hin. Wie aus Abbildung 28b hervorgeht, ergeben sich für die Fixationsdauer mit einer Ausnahme positive Korrelationen. Auch hier weisen die meisten Korrelationen auf mittlere bis starke Zusammenhänge hin. Hinsichtlich EEG-Frustration ergeben sich auf Individualebene zumeist negative, teils aber auch positive

Korrelationen mit der Leistung (Abbildung 28d). Im Vergleich zu den Individualkorrelationen der anderen Maße fallen diese deutlich schwächer aus.

Insgesamt sprechen auch die Ergebnisse auf Individualebene für die Bestätigung der in H3b formulierten Hypothesen zum Zusammenhang zwischen dem *Punktestand* und den Maßen *Pupillenweite, Fixationsdauer* und *EEG-Engagement*. Jedoch ist zu beachten, dass die Höhe der Zusammenhänge interindividuell unterschiedlich ausfällt. Für *EEG-Frustration* lässt sich Hypothese H3b aufgrund der geringen Zusammenhänge mit dem Punktestand auf Gruppen- und auf individueller Ebene hingegen nicht bestätigen.

Post hoc-Untersuchung: Kombination mehrerer physiologischer Maße

Wie sich bei der Analyse der Zusammenhänge auf Individualebene herausstellte, korrelieren die physiologischen und verhaltensbasierten Maße bei den Versuchsteilnehmern unterschiedlich stark mit der subjektiven Bewertung und der Leistung. Um eine robustere Erfassung des Nutzerzustands auf individueller Ebene zu gewährleisten, wird in der Literatur empfohlen, mehrere Einzelmaße zu kombinieren (vgl. Abschnitt 2.4.2). Neben der hypothesenprüfenden Untersuchung wurde daher auf explorativer Basis zusätzlich untersucht, ob der Zusammenhang mit der Leistung auf Individualebene durch eine Kombination der Einzelmaße *EEG-Engagement*, *Pupillenweite* und *Fixationsdauer* verbessert werden kann. Hierzu wurden für jede Person multiple Regressionen berechnet, in die die drei Indikatoren als Prädiktorvariablen einbezogen werden und der *Punktestand* die Kriteriumsvariable darstellt. Die Ergebnisse der multiplen Regressionen sind in Tabelle 28 aufgeführt und den Ergebnissen der bivariaten Regressionen gegenübergestellt, die mit dem jeweiligen besten Prädiktor durchgeführt wurden.

Tabelle 28. Ergebnisse der regressionsanalytischen Untersuchungen zur Vorhersage der Leistung durch einzelne und kombinierte Eyetracking- und EEG-Maße auf Individualebene (Experiment 1)

		Bivariate Regression				Multiple Regression			
VP	**N**	**Bester Prädiktor**	**β**	**R^2**	***F* (1,6)**	**R**	**R^2**	***F* (3,4)**	**Δ in R^2**
5	8	Pupillenweite	-.84*	.7	13,88	.94	.89*	10,99	.19
2	8	Fixationsdauer	.89*	.79	22,5	.91	.83	6,40	.04
3	8	Pupillenweite	-.88*	.78	20,94	.89	.8	5,23	.02
4	8	Engagement	-.73*	.53	6,84	.86	.75	2,93	.22
6	7	Pupillenweite	-.81	.66	11,78	.87	.76	4,27	.10
8	8	Fixationsdauer	.69	.47	5,32	.73	.54	1,54	.07
7	7	Engagement	-.37	.14	0,95	.59	.35	0,82	.21
10	8	Engagement	-.54	.29	2,43	.54	.29	0,56	0
14	8	Fixationsdauer	.42	.17	1,27	.48	.23	0,76	.05
9	8	Fixationsdauer	.23	.05	0,34	.31	.1	0,27	.05

Anmerkungen: Die Tabelle ist sortiert nach dem Anteil erklärter Varianz (R^2) in der multiplen Regression. Bei VP09 und VP14 erfolgte die multiple Regression aufgrund mehrerer fehlender Werte ohne die Variable Engagement. $^{**}p<.01$, $^{*}p<.05$; bei allen Regressionen ist die Toleranz TOL >.3, es liegt somit keine Multikollinearität vor (vgl. Field, 2009, Bortz, 2005).

Die Ergebnisse zeigen, dass nur für VP 5 eine signifikante Verbesserung der Vorhersageleistung durch die multiple Regression erzielt werden kann. Allerdings ist bei der Beurteilung der Signifikanzen die sehr geringe Fallzahl zu berücksichtigen. Die Nützlichkeit U kann anhand der

Änderung in R^2 ermittelt werden, die sich bei Hinzunahme der weiteren Prädiktoren ergibt (vgl. Bortz, 2005). Dabei zeigt sich, dass die multiple Regression bei den Probanden VP 4, VP 5 und VP 7 mit einem Zuwachs in R^2 um circa .20 verbunden ist und somit den Anteil der erklärten Varianz deutlich erhöht. Dies spricht dafür, dass eine Kombination der Maße die Güte der Vorhersage zumindest bei einigen Personen verbessern kann.

H3c: Zusammenhänge zwischen Leistungsmaßen und subjektiven Maßen

Zuletzt wird untersucht, inwiefern die subjektiven Maße mit dem Punktestand als Leistungsmaß korrelieren. Tabelle 29 gibt die Korrelationen für den NASA-TLX-Gesamtscore, die Subskala Frustration sowie die Dimensionen Valenz und Arousal des SAM mit dem Punktestand auf Gruppenebene wieder. In den Spalten r_{part} sind, wie zuvor, die Partialkorrelationen aufgeführt, bei denen etwaige Zwischensubjektunterschiede auspartialisiert wurden. Abbildung 29 veranschaulicht die Individualkorrelationen zwischen dem NASA-TLX-Gesamtscore und dem Punktestand als Balkendiagramme.

Tabelle 29. Produkt-Moment-Korrelationen zwischen dem Punktestand und den subjektiven Maßen (NASA-TLX und SAM) auf Gruppenebene unkorrigiert (r) und nach Bereinigung von Zwischensubjekteffekten (r_{part}) – Experiment 1

	NASA-TLX-Gesamtscore		NASA-TLX Frustration		Valenz SAM		Arousal SAM	
	r	r_{part}	*r*	r_{part}	*r*	r_{part}	*r*	r_{part}
Punktestand Fallzahl: 96	-.65**	-.75**	-.50**	-.61**	.39**	.52**	-.36**	-.51**

** p<.01; * p<.05 bei einseitigem Testen

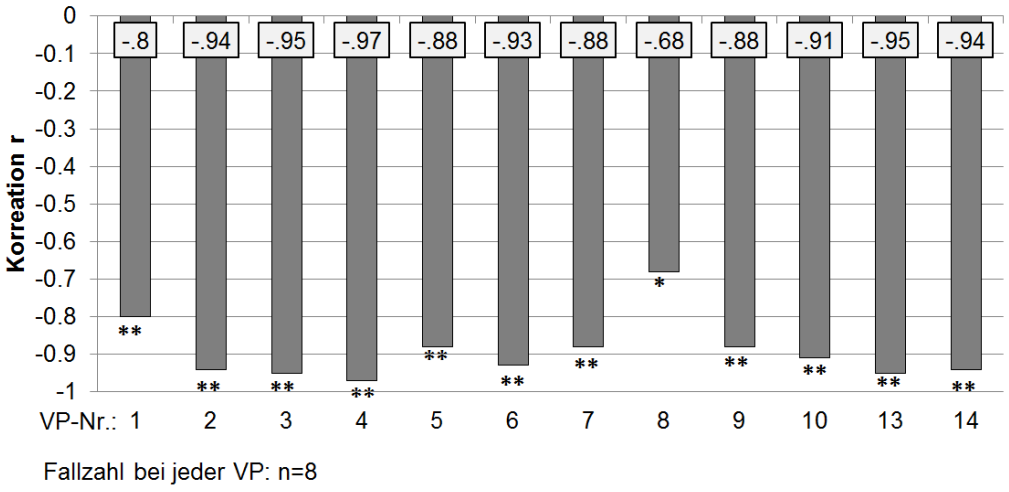


Abbildung 29. Individualkorrelationen pro Versuchsperson (VP) zwischen dem Punktestand und dem NASA-TLX-Gesamtscore (p<.01, *p<.05)**

Alle Maße weisen auf Gruppenebene jeweils signifikante Korrelationen (p<.01) mit dem Punktestand auf. Die negative Ausrichtung der Korrelationen bei den NASA-TLX-Maßen spricht dafür, dass die Leistung, wie erwartet, schlechter ausfällt, je höher die Beanspruchung und Frustration ist. Bei den SAM-Dimensionen weist die positive Korrelation für Valenz und die negative Korrelation für Arousal darauf hin, dass eine positive Stimmung und ein geringerer

Erregungsgrad mit einer besseren Leistung einhergehen. Die Partialkorrelationen weisen durchweg auf höhere Zusammenhänge hin als die unbereinigten Korrelationen.

Auf Individualebene fallen die Zusammenhänge größtenteils noch höher aus. Bei nahezu allen Versuchspersonen ergeben sich fast perfekte Korrelationen mit dem Punktestand. Insgesamt können die Hypothesen zu H3c zum Zusammenhang zwischen Leistungsmaßen und subjektiven Maßen auf Gruppen- und insbesondere auch auf Individualebene bestätigt werden.

4.4 Diskussion

Abschließend werden die Ergebnisse aus dem Experiment noch einmal zusammenfassend dargestellt und diskutiert. Es werden Schlussfolgerungen für die Diagnose des multidimensionalen Nutzerzustands gezogen, und es wird ein Überblick gegeben, welche Anforderungen aus den Kapiteln 2 und 3 im Experiment umgesetzt wurden, und welche Erkenntnisse aus der Umsetzung gewonnen werden können.

4.4.1 Diskussion der Ergebnisse zu Hypothese 1

Tabelle 30 bietet einen Überblick, welche der in Abschnitt 4.2.7 formulierten statistischen Hypothesen zu im Rahmen der inferenzstatistischen Auswertung (Abschnitt 4.3.1) bestätigt bzw. nicht bestätigt werden konnten. Die Befunde weisen darauf hin, dass die mentale Beanspruchung durch die Variation der Anzahl zu überwachender Areas erfolgreich moduliert werden konnte. Dies zeigt sich darin, dass alle Beanspruchungsindikatoren hypothesenkonform signifikante Veränderungen in erwarteter Richtung zwischen den Bedingungen mit 2 und mit 5 Areas aufweisen.

Tabelle 30. Ergebnisübersicht zum Einfluss der Anforderungsmerkmale auf den diagnostizierten Nutzerzustand (Hypothese 1 – Experiment 1)

		Veränderung durch Anzahl Areas	Veränderung durch Kooperativität	Veränderung durch Lärm
H1a	NASA-TLX-Gesamt	ja	ja	nein
	NASA-TLX-Frustration	ja	ja	nein
	SAM Valenz	ja	nein	nein
	SAM Arousal	ja	nein	nein
H1b	Pupillenweite	ja	nein	nicht auswertbar
	Fixationsdauer	ja	nein	nein
	EEG-Engagement	ja	nein	nein
	EEG-Frustration	nein	nein	ja
H1c	Punktestand	ja	ja	nein
	Ausg. Rechenaufgaben	ja	ja	ja
	Dauer der Alarme	ja	ja	ja

grün: hypothesenkonform; rot: nicht hypothesenkonform; hellgrau: keine Hypothese angenommen

Zusätzlich wirkte sich der Faktor *Anzahl Areas* auch auf die Frustration und die Leistung aus, wie sich an ebenfalls signifikanten Veränderungen durch diesen Faktor bei den subjektiven Maßen zur Frustration und den Leistungsmaßen zeigt. Zu den Auswirkungen der Anzahl Areas auf die Frustration wurde keine Hypothese formuliert. Eine mögliche Erkärung für die Ergebnisse könnte sein, dass die längere Dauer von Alarmen und die häufigeren Regelverletzungen in der Bedingung mit 5 Areas zu einer höheren Frustration der Teilnehmer geführt haben, da sie instruiert worden waren, Regelverletzungen (die sich durch Alarme ankündigten) zu vermeiden. Außerdem bemerkten einige Teilnehmer in der abschließenden mündlichen Befragung, dass sie den Ton der Alarme als unangenehm empfunden haben. Inwiefern sich Beanspruchung, Leistung und Frustration tatsächlich kausal beeinflusst haben, kann jedoch im Rahmen der Analysen nicht geklärt werden.

Der Faktor *Kooperativität*, der zur Modulierung der Frustration herangezogen wurde, weist gegenüber dem Faktor Anzahl Areas weniger deutliche Effekte auf. Eine signifikante Veränderung des emotionalen Zustands durch den Faktor Kooperativität kann zwar durch die NASA-TLX-Skala Frustration, nicht jedoch durch die Dimensionen des SAM und den EEG-Klassifikator Frustration nachgewiesen werden. Dies deutet darauf hin, dass sich der emotionale Zustand nur mäßig durch die unterschiedlich hohe Kooperativität der Kontakte verändert hat. Es zeigt aber auch, dass die NASA-TLX-Skala gegenüber dem SAM und dem EEG-Klassifikator eine höhere Sensitivität für die Erfassung von Unterschieden in der Frustration aufzuweisen scheint.

Trotz der schwächeren Auswirkungen der Kooperativität der Kontakte auf die Frustration zeigt sich, dass dieser Faktor signifikanten Einfluss auf die Leistung genommen hat. Dabei ist zu beachten, dass neben einer geringeren Leistung in der Bedingung mit geringer Kooperativität die subjektive Beanspruchung signifikant höher bewertet wurde als in der Bedingung mit hoher Kooperativität. Die geringe Kooperativität der Kontakte könnte die Aufgabe daher erschwert und sich so auf die Leistung ausgewirkt haben.

In Bezug auf den Faktor *Lärm* wurde angenommen, dass das Vorhandensein von Lärm das Arousal erhöht und den Fokus der Aufmerksamkeit einschränkt. Dies sollte sich in einem höheren Engagement und einer schlechteren Leistung in der Bedingung mit Lärm äußern. Für das *EEG-Engagement* konnte dies jedoch nicht bestätigt werden. Bei der Leistung bestätigt sich die Annahme nur in Hinblick auf die Leistungsmaße *Ausgelassene Rechenaufgaben* und *Dauer der Alarme*. Ein unerwartetes Ergebnis bei dem Faktor *Lärm* zeigte sich außerdem in Bezug auf den EEG-Klassifikator *Frustration*. Die Analysen weisen darauf hin, dass die Bedingung mit Lärm mit einer geringeren Frustration verbunden war als die Bedingung ohne Lärm. Ein solcher Zusammenhang wird jedoch nicht durch die subjektiven Maße gestützt. Da der Lärm einen Störfaktor darstellen sollte, wäre außerdem eher zu erwarten, dass die Frustration in der Lärmbedingung höher liegt. Die Validität des Frustrationsklassifikators muss in Hinblick auf diesen Befund daher in Frage gestellt werden.

Zusammenfassend kann konstatiert werden, dass die Hypothesen H1a-H1c für den Faktor *Anzahl Areas* bestätigt werden können. Für die Faktoren *Kooperativität* und *Lärm* trifft dies hingegen, wie Tabelle 30 verdeutlicht, nur teilweise zu.

Die Analyse und Diskussion der Ergebnisse führte darüber hinaus zu den Erkenntnissen,

- dass sich die betrachteten Merkmale der Anforderungssituation unterschiedlich stark auf den Nutzerzustand auswirken,
- dass sich ein Anforderungsmerkmal auch auf verschiedene Nutzerzustandsdimensionen gleichzeitig auswirken kann,
- und dass sich Veränderungen des Nutzerzustands nicht zwingend bei allen Diagnosemaßen gleichermaßen zeigen.

4.4.2 Diskussion der Ergebnisse zu Hypothese 2

In Hypothese 2 wurde angenommen, dass die individuellen Faktoren Fähigkeiten, Erfahrung, Müdigkeit und Motivation interindividuelle Unterschiede in der Beanspruchung, dem Arousal und der Leistung erklären können. Eine Voraussetzung ist, dass zwischen diesen Variablen Korrelationen in erwarteter Richtung bestehen. Wie Tabelle 31 im Überblick zeigt, weisen nahezu alle individuellen Faktoren Korrelationen in erwarteter Richtung mit den Indikatoren des Nutzerzustands (NASA-TLX und Arousal) sowie den Leistungsmaßen auf.

Tabelle 31. Ergebnisübersicht zu den Korrelationen zwischen individuellen Faktoren und den Nutzerzustands- und Leistungsmaßen (Hypothese 2 – Experiment 1)

		L-R-Test	Rechentest	Erfahrung	Müdigkeit	Motivation
H2a	NASA-TLX	negativ	unkorreliert	negativ		positiv
	Arousal (SAM)				negativ	positiv
H2c	Dauer Alarme	unkorreliert	negativ	negativ	positiv	positiv
	Ausgelassene Rechenaufgaben	negativ	negativ	negativ	positiv	negativ
	Punktestand	positiv	positiv	positiv	negativ	positiv

negativ = negative Korrelation, positiv = positive Korrelation; unkorreliert = $|r|<.05$;
grün = hypothesenkonform; hellgrün = hypothesenkonform aber nicht signifikant; rot = nicht hypothesenkonform; hellgrau = keine Hypothese angenommen

Als signifikant erwiesen sich hauptsächlich die Zusammenhänge mit den Leistungsmaßen, die mit Korrelationen im Bereich $|r|>.3$ als mittel bis hoch einzustufen sind. Eine Ausnahme bildet die (schlafbezogene) Müdigkeit, die zwar erwartungskonforme aber nicht signifikante Zusammenhänge mit der Leistung aufweist. Literaturbefunde legen hingegen nahe, dass Müdigkeit die Leistung in starkem Maße negativ beeinflusst (vgl. Abschnitt 3.1.4). Eine Erklärung findet sich bei Betrachtung der Müdigkeitsbewertungen in beiden Durchgängen. Hierbei zeigt sich, dass in 19 von 24 Fällen die zweite von sieben Abstufungen der Müdigkeit gewählt wurde. Die Müdigkeit war bei den meisten Teilnehmern somit nur schwach ausgeprägt und hat sich kaum interindividuell unterschieden. Dementsprechend konnte sich dieser Einflussfaktor nicht bedeutsam auf die Leistung auswirken.

Die signifikanten Korrelationen sagen jedoch noch nicht aus, ob die Zusammenhänge kausal interpretiert werden können. Da die individuellen Faktoren zu Beginn des Experiments und somit zeitlich vor den Leistungsmaßen erhoben wurden, ist eine gegenläufige Ursache-Wirkung-Relation

als Erklärung für die Zusammenhänge unplausibel (vgl. Bortz & Döring, 2006, S. 519). Es ist allerdings zu beachten, dass zwischen den individuellen Faktoren und den Leistungsmaßen nicht zwangsläufig direkte kausale Zusammenhänge bestehen müssen. So könnten die Zusammenhänge auch durch Moderatorvariablen zustande gekommen sein, die sich auf beide Faktoren ausgewirkt haben. Zum Beispiel wäre für die signifikant negative Korrelation zwischen dem *Rechentest* und der *Dauer der Alarme* denkbar, dass der Zusammenhang durch eine nicht betrachtete Moderatorvariable, wie die Intelligenz, hervorgerufen wurde. Intelligente Menschen würden dann – sofern die These richtig ist – eine bessere Leistung im Rechentest als auch eine bessere Leistung im Versuch erzielen (z.B. durch das Entwickeln von Strategien zur Reduzierung der Belastung, vgl. Abschnitt 4.5.2). Da die Intelligenz nicht erfasst wurde, kann diese These nicht überprüft werden. Dennoch zeigt es, dass Alternativhypothesen zu kausalen Zusammenhängen nicht ausgeschlossen werden können.

Es sei auch angemerkt, dass während der Versuchsdurchführung auch noch ein weiterer potenziell bedeutsamer Einflussfaktor identifiziert werden konnte. So zeigte sich, dass Teilnehmer mit schlechten Englischkenntnissen Probleme hatten, die Rechenaufgaben zu lösen, da diese auf Englisch gestellt wurden. Insofern ist zu vermuten, dass sich zwischen Englischkenntnissen und ausgelassenen Rechenaufgaben ein negativer Zusammenhang zeigen würde.

Insgesamt kann festgehalten werden, dass eine Vielzahl an individuellen Eigenschaften und Merkmalen existiert, die sich bedeutsam auf Nutzerzustand und Leistung auswirken können. Hypothese 2 kann somit auf allgemeiner Ebene als bestätigt betrachtet werden. Allerdings erscheint es notwendig, Abhängigkeiten zwischen den jeweiligen individuellen Faktoren und möglichen nicht berücksichtigten Einflussgrößen zu prüfen, bevor sie zur Identifizierung von Ursachen für kritische Nutzerzustände und Leistungseinbußen herangezogen werden.

4.4.3 Diskussion der Ergebnisse zu Hypothese 3

In Hypothese 3 wurde angenommen, dass die Diagnosemaße korrelieren, die Literaturbefunden zufolge den gleichen Nutzerzustand erfassen. Die wesentlichen Ergebnisse, die sich für die statistischen Hypothesen zu H3a-H3c ergeben haben, sind in Tabelle 32 zusammengefasst.

Tabelle 32. Ergebnisübersicht zu den Korrelationen zwischen den Diagnosemaßen (Hypothese 3 – Experiment 1)

		Pupillenweite	Fixationsdauer	EEG-Engagement	EEG-Frustration	**H3c** Punktestand
H3a	NASA-TLX-Gesamt	positiv	negativ	positiv		negativ
	NASA-TLX Frustration				positiv	negativ
	SAM Valenz				unkorreliert	positiv
	SAM Arousal			positiv	negativ	negativ
H3b	Punktestand	negativ	positiv	negativ	negativ	1

negativ = negative Korrelation, positiv = positive Korrelation; unkorreliert = |r|<.05;
grün = hypothesenkonform, hellgrün = hypothesenkonform aber nicht signifikant (auf Gruppenebene), rot = nicht hypothesenkonform, hellgrau = keine Hypothese angenommen

Im Fokus stehen insbesondere die Ergebnisse in Bezug auf die physiologischen und verhaltensbasierten Maße, da sie sich in der theoretischen Analyse als vielversprechend für eine Diagnose des Nutzerzustands in adaptiven Systemen herausgestellt haben (vgl. Abschnitt 2.3). Mit

Ausnahme des Klassifikators *EEG-Frustration* korrelieren diese in erwarteter Richtung und zumeist auch signifikant mit den subjektiven Maßen und den Leistungsmaßen. Die Ergebnisse bestätigen somit die Validität dieser Maße.

Für *EEG-Frustration* kann die Validität hingegen nicht bestätigt werden. Dieser Klassifikator korreliert mit der NASA-TLX-Skala *Frustration* zwar in erwarteter Richtung, aber nur schwach und nicht signifikant. Mit den SAM-Dimensionen *Valenz* und *Arousal* kann keine Korrelation in erwarteter Richtung festgestellt werden. Auch andere Forscher gelangten zu einem ähnlichen Ergebnis. So berichten auch Wright (2010) sowie Noronha, Sol, & Vourvopoulos (2013), dass der EEG-Klassifikator *Frustration* in ihren Studien nicht mit Selbstberichtsangaben korrelierte. Es kann somit gefolgert werden, dass der Klassifikator *EEG-Frustration* kein geeigneter Indikator ist, um den emotionalen Zustand der Frustration zu erfassen.

Die Betrachtung der Zusammenhänge auf Individualebene machte außerdem deutlich, dass zwischen den Personen teilweise starke Unterschiede darin bestehen, wie hoch die einzelnen Maße mit der subjektiven Beanspruchung und der Leistung korrelieren. Wie die Post hoc-Untersuchung verdeutlichte, kann für die meisten Versuchsteilnehmer ein Indikator identifiziert werden, der hoch mit der Leistung korreliert (vgl. Tabelle 28 bivariate Regression). Dies spricht für die in Abschnitt 2.4.3 formulierte Forderung, dass die Auswahl geeigneter Diagnosemaße bei einer individuellen Echtzeitdiagnose personenspezifisch erfolgen sollte. Allerdings konnten im Vorhinein keine Hypothesen formuliert werden, welche Maße für welche Personen am besten geeignet sein würden. Die identifizierten Unterschiede könnten daher auch zufällig zustande gekommen sein. Inwiefern sich eine personenspezifische Eignung bestätigen lässt, wird im Rahmen des Retests in Kapitel 5 untersucht.

Die Ergebnisse der multiplen Regression zeigten außerdem, dass durch die Kombination von EEG-Engagement, Pupillenweite und Fixationsdauer die Vorhersagequalität teilweise deutlich verbessert werden kann. Dies spricht für die in Abschnitt 2.4.2 identifizierte Anforderung, dass verschiedene Diagnosemaße kombiniert werden sollten.

4.5 Resümee

4.5.1 Schlussfolgerungen für die Erfassung des multidimensionalen Nutzerzustands

Insgesamt weisen die Ergebnisse der experimentellen Untersuchung darauf hin, dass eine Bestimmung von Unterschieden in der mentalen Beanspruchung durch die verwendeten Maße gut möglich ist. Dabei ist allerdings zu berücksichtigen, dass die Analysen auf Basis der gemittelten Werte pro Test durchgeführt wurden. Es bleibt somit zu untersuchen, inwieweit diese Maße auch in der Lage sind, Veränderungen des Nutzerzustands in Echtzeit anzuzeigen.

Für die Diagnose des emotionalen Zustands und der Aufmerksamkeit erscheint es angesichts der Ergebnisse sinnvoll, andere Maße in die Diagnose einzubeziehen. Der emotionale Zustand kann zwar zufriedenstellend über Fragebögen erfasst werden, allerdings sind subjektive Maße für den Einsatz in adaptiven Systemen, wie in Abschnitt 2.3.2 dargelegt wurde, weniger geeignet. Eine physiologische Erfassung über den EEG-Klassifikator *Frustration* stellte sich dagegen als nicht valide heraus. Zur Erfassung der „Arousal“-Komponente des emotionalen Zustands könnte zum

Beispiel die Herzrate herangezogen werden. Für die Bestimmung der Valenz könnten EMG-Klassifikatoren verwendet werden, die eine Detektion des Gesichtsausdrucks ermöglichen (vgl. Abschnitt 2.2.2).

Bei der Aufmerksamkeit zeigte sich die Herausforderung, dass diese nicht mit bestehenden subjektiven Bewertungsverfahren erfasst werden kann, so dass in der experimentellen Untersuchung kein geeignetes Vergleichsmaß herangezogen werden konnte, um den EEG-Klassifikator *Engagement* als Diagnosemaß für die Aufmerksamkeit zu validieren. Denkbar wäre in künftigen Untersuchungen über die Methoden *SAGAT*, *SPAM* oder *SART* (vgl. Abschnitt 2.2.6) Rückschlüsse auf die Aufmerksamkeit zu ziehen, da die erste Stufe des Situationsbewusstseins (Level-1-SA) nach der Definition von Endsley (1988) eng mit der Aufmerksamkeit verknüpft ist (vgl. Abschnitt 3.2).

4.5.2 Erkenntnisse zu den Anforderungen aus dem Stand der Forschung

In Experiment 1 lag der Fokus darauf, Einflussfaktoren auf den Nutzerzustand und Auswirkungen des Nutzerzustands zu untersuchen, die bei der Analyse psychologischer Modelle und Theorien in Kapitel 3 identifiziert worden waren. Die wesentlichen Erkenntnisse dazu wurden in den vorigen Abschnitten bereits zusammengefasst und bewertet. Darüber hinaus adressiert das Experiment jedoch auch die Anforderungen und Empfehlungen, die aus dem Stand der Forschung zur Nutzerzustandsdiagnose im Kontext adaptiver Systeme abgeleitet wurden (vgl. Abschnitt 2.4). Tabelle 33 zeigt eine Übersicht, wie die Anforderungen, die in Abschnitt 2.4 identifiziert worden waren, im Experiment umgesetzt oder berücksichtigt wurden. Außerdem ist dargestellt, welche Ergebnisse und Erkenntnisse diesbezüglich durch die experimentelle Umsetzung gewonnen werden konnten. Auf diese wurde teilweise bereits bei der Diskussion der Experimentalergebnisse eingegangen, z.B. in Hinblick auf die individuelle Betrachtung und die Kombination von verschiedenen Maßen.

Insgesamt belegen die Experimentalergebnisse die Sinnhaftigkeit der Empfehlungen und Anforderungen. In Bezug auf die Forderung, dass realitätsnahe Aufgaben verwendet werden sollten, wurden im Experiment allerdings auch nachteilige Aspekte deutlich. So zeigte sich, dass die verwendete Experimentalaufgabe zu einer geringeren Standardisierung und weniger vergleichbaren Versuchsbedingungen führte, da die Teilnehmer durch ihre Nutzerinteraktion selbst den Verlauf eines Tests und die damit einhergehenden Belastungen beeinflussen konnten. Zum Beispiel wandten einzelne Teilnehmer die Strategie an, Kontakte 360° um ihre eigene Achse drehen zu lassen. Dies führte dazu, dass die Kontakte für die Zeit der Drehung an etwa der gleicher Stelle blieben, so dass die Gefahr für Grenzüberschreitungen und Kollisionen reduziert war. Außerdem mussten die Teilnehmer dadurch weniger häufig Richtungsanweisungen geben, was die Belastung reduzierte. Diese Strategie hat somit vermutlich zu interindividuellen Unterschieden im Nutzerzustand und der Leistung beigetragen. Da das Steuerverhalten jedoch nicht erfasst und als Variable berücksichtigt wurde, handelt es sich hierbei um einen Störeinfluss, der sich in Hinblick auf die statistische Auswertung in einer höheren Fehlervarianz äußert. Es scheint somit für die systematische Untersuchung von Effekten sinnvoll, Experimentalaufgaben zu verwenden, die trotz Realitätsnähe ein gewisses Maß an Standardisierung ermöglichen (vgl. Abschnitt 6.3).

Tabelle 33. Erkenntnisse zu den in Experiment 1 berücksichtigten Anforderungen aus Abschnitt 2.4

Umsetzung der Anforderung	Erkenntnisse
1. **Störeinflüsse sollten mit erhoben oder konstant gehalten werden.**	
• Helligkeit wurde konstant gehalten. • Tageszeit wurde intraindividuell in beiden Durchgängen konstant gehalten. • Müdigkeit und Motivation wurden durch Befragung erfasst.	• Auf den Nutzerzustand und die Leistung wirken sich vielfältige Faktoren aus, wobei nicht alle experimentell erfasst wurden (z.B. Intelligenz, Englischkenntnisse, vgl. Abschnitt 4.4.2)
2. **Verschiedene Methoden zur Erfassung des Nutzerzustands sollten kombiniert werden.**	
• Kombination von physiologischen/ verhaltensbasierten Maßen in multipler Regression	• Kombination von Fixationsdauer, Pupillenweite und Engagement ermöglicht bei den meisten Probanden eine bessere Vorhersage der Leistung als die Einzelmaße.
3. **Physiologische Sensoren sollten an einer Baseline relativiert werden; Indikatoren sollten auf Individualebene in Hinblick auf ihre Eignung geprüft und nutzerspezifisch für die Diagnose ausgewählt werden.**	
• Relativierung der Messwerte erfolgte durch z-Transformation. • Untersuchung der Zusammenhänge auf Individualebene.	• Maße korrelieren interindividuell unterschiedlich stark → Untersuchung erforderlich, ob personenspezifische Effekte zeitlich stabil sind (siehe Kapitel 5).
4. **Zustände, in denen der Operateur sich nicht mehr selbst adaptieren kann, sollten diagnostiziert werden.**	
• Regelverletzungen und Auslassungen von Rechenaufgaben weisen im Experiment auf Leistungsprobleme hin.	• Hohe Korrelationen zwischen den Leistungsmaßen und den Indikatoren des Nutzerzustands weisen darauf hin, dass Leistungsprobleme durch „kritische“ Nutzerzustände hervorgerufen wurden, in denen die Selbstregulierung nicht erfolgreich war.
5. **Die Diagnose sollte bereits bei den Ursachen für kritische Nutzerzustände einsetzen.**	
• Anforderungsmerkmale wurden zur Modulierung kritischer Nutzerzustände herangezogen. • Erfassung von individuellen Faktoren.	• Untersuchte Faktoren kommen als mögliche Ursachen für Unterschiede in Nutzerzustand und Leistung in Betracht. Kausale Abhängigkeiten können jedoch nicht nachgewiesen werden.
6. **Kontextinformationen und Aufgabenstatus sollten identifiziert und erfasst werden.**	
• verschiedene aufgabenspezifische Ereignisse (z.B. Alarme, Kollisionen) wurden mitgeloggt.	• aufgabenspezifische Ereignisse können Leistungsminderungen anzeigen.
7. **Es sollten Verfahren zur „Glättung“ der physiologischen Daten angewendet werden.**	
• Bei Betrachtung von Unterschieden zwischen Versuchsbedingungen nicht erforderlich.	
8. **Bei der experimentellen Untersuchung sollten realitätsnahe Aufgaben verwendet werden (z.B. Verwendung von Simulatoren).**	
• Verwendung eines Demonstrators, der in Funktionalität und Design auf einem realen Trainingssystem für Aufgaben der Luftraumüberwachung basiert	• Szenario mit gleichen Ausgangsbedingungen kann sich bei jedem Teilnehmer anders entwickeln, da die Nutzerinteraktion den weiteren Verlauf beeinflusst. → Problem: Standardisierung nicht gegeben, Erhöhung der Fehlervarianz

5 Experiment 2 – Retest zur Untersuchung der zeitlichen Stabiltät

Das in Kapitel 4 beschriebene Experiment zur multifaktoriellen Nutzerzustandsdiagnose wurde ein Jahr später mit 10 Personen, die an dem Experiment teilgenommen hatten, wiederholt. Durch den Retest sollten zwei Forschungsziele untersucht werden: zum Einen sollte geprüft werden, ob die untersuchten Eyetracking- und EEG-Maße eine ausreichende zeitliche Stabilität aufweisen; das heißt, ob sich die Ergebnisse aus dem ersten Experiment durch den Retest replizieren lassen. Zum Anderen sollte untersucht werden, inwiefern sich weitere physiologische und verhaltensbasierte Maße, die über den Brustgurt *Zephyr BioHarness 3* erfasst wurden, als Indikatoren des Nutzerzustands eignen. Auf beide Forschungsziele wird im Folgenden näher eingegangen.

5.1 Forschungsziele

5.1.1 Untersuchung der zeitlichen Stabilität der Eyetracking- und EEG-Maße

Während sich eine Vielzahl an Studien mit der Validität und Sensitivität von physiologischen Maßen für die Erfassung von Nutzerzuständen beschäftigt, wurde ihre Reliabiliät und ihre zeitliche Stabilität[5] deutlich seltener empirisch untersucht. Faulstich et al. (1986) führten eine Studie durch, in der sie die zeitliche Stabilität mehrerer physiologischer Maße (u.a. Körpertemperatur, Herzrate, Blutdruck, Elektromyografie) in Hinblick auf ihr Verhalten gegenüber physischen und mentalen Stressoren untersuchten. Der Retest wurde zwei Wochen nach dem ersten Test durchgeführt. Faulstich et al. (1986) fanden heraus, dass die meisten physiologischen Maße moderate bis hohe Test-Retest-Korrelationen aufwiesen (zumeist im Bereich zwischen r=.4 und r=.8), wenn absolute Baseline und Testwerte zugrunde gelegt wurden. Allerdings zeigte sich, dass die an der Baseline relativierten Werte keine adäquate Reliabilität aufwiesen.

Tomarken (1995) bemerkt in seinem Review zu physiologischen Messmethoden, dass physiologische Maße in den von ihm betrachteten Studien bedeutsame Unterschiede in der zeitlichen Stabilität aufwiesen. Die Test-Retest-Korrelationen würden dabei in der Regel zwischen r=.3 und r=.6 schwanken. Ähnliche Unterschiede zeigten sich auch in der Studie von Faulstich et al. (1986). Zu beachten ist, dass der zeitliche Abstand zwischen den Erhebungszeitpunkten in den Studien nur wenige Tage oder Wochen beträgt, während der zeitliche Abstand im vorliegenden Retest-Experiment mit einem Jahr bedeutend höher ausfällt. Es gibt kaum Studien, welche die zeitliche Stabilität von physiologischen Maßen über diesen Zeitraum hinweg untersucht haben. Eine Untersuchung wurde von Uhlig (2018) für die HRV vorgenommen. Darin zeigte sich eine hohe unsystematische Variabilität der Messwerte zwischen den Erhebungszeitpunkten. Die Erfassung erfolgte dabei im Ruhezustand und bekannte Störeinflüsse wurden kontrolliert oder systematisch

[5] Nach Tomarken (1995) liegt zeitliche Stabilität vor, wenn sich Befunde bei einem zeitlich später stattfindenden Retest replizieren lassen, während Reliabilität anzeigt, inwiefern Befunde bei mehrmaliger Messung innerhalb einer Studie konsistent sind.

variiert (Körperposition). Die HRV kann in der vorliegenden Untersuchung zwar nicht in Hinblick auf ihre zeitliche Stabilität bewertet werden. Sie stellt jedoch einen potenziellen Indikator für die mentale Beanspruchung dar, der über den BioHarness erfasst werden kann (vgl. Abschnitt 5.2.3).

Zu beachten ist, dass die Analysen zur zeitlichen Stabilität vorwiegend auf Gruppenebene durchgeführt wurden. Für die eigene Fragestellung ist es hingegen von vorrangigem Interesse, ob die verwendeten Maße eine zeitlich stabile Diagnose des Nutzerzustands auf individueller Ebene ermöglichen. Daher erscheint es notwendig die zeitliche Stabilität der verwendeten Diagnosemaße auch auf Individualebene zu untersuchen. Insbesondere soll geprüft werden, ob die interindividuell unterschiedliche Sensitivität der Maße, die im ersten Experiment beobachtet wurde (vgl. Abschnitt 4.3.3), zeitlich stabil ist. Da im Vorhinein keine konkreten Annahmen zu den Ergebnissen auf Individualebene in Experiment 1 gemacht werden konnten, handelt es sich dabei um explorative Befunde. Die Annahme einer personenspezifischen Eignung der Diagnosemaße (vgl. Abschnitt 2.4.3) lässt sich somit erst bestätigen, wenn sich die Befunde im Retest replizieren lassen.

5.1.2 Untersuchung der diagnostischen Fähigkeiten des Zephyr BioHarness 3

Im Retest wurde der Zephyr BioHarness 3 (kurz: BioHarness) als weiterer Sensor hinzugenommen (vgl. Sensorbeschreibung in Abschnitt 5.2.2). Da sich im vorangegangenen Experiment herausstellte, dass die Frustration über den Frustrationsklassifikator des Emotiv EEG nicht valide erfassbar ist, wurde es als notwendig erachtet, weitere Maße zur Diagnose des emotionalen Zustands heranzuziehen. Nach Russell (1980) sind emotionale Zustände durch die Valenz (positiv vs. negativ) und das Arousal (ruhig vs. erregt) gekennzeichnet. Der BioHarness erfasst eine Vielzahl an physiologischen Parametern, die mit dem Arousal in Zusammenhang stehen (z.B. Herzrate; HRV und Atemfrequenz), so dass die Diagnose des emotionalen Zustands hierdurch verbessert werden könnte. Viele Parameter, die durch den BioHarness erfasst werden, stehen außerdem mit der Beanspruchung und möglicherweise weiteren Nutzerzustandsdimensionen in Zusammenhang. Durch das Retest-Experiment sollen die diagnostischen Fähigkeiten der durch den BioHarness gewonnenen Maße daher im Vergleich zu den bereits vorhandenen Eyetracking- und EEG-Maßen untersucht und bewertet werden.

5.2 Methodisches Vorgehen

Da es sich um einen Retest handelt, stimmt das methodische Vorgehen weitgehend mit dem des ersten Experiments überein (vgl. Abschnitt 4.2). Im Folgenden wird daher vorwiegend auf die Aspekte näher eingegangen, die sich im Vergleich zur ersten Untersuchung unterscheiden.

5.2.1 Versuchsdesign

Bei der Stichprobe (N= 10) handelt es sich um acht männliche und zwei weibliche Mitarbeiter des Fraunhofer FKIE im Alter zwischen 20 und 39 Jahren (Ø 31,4 Jahre), die ein Jahr zuvor an dem ersten Experiment (Kapitel 4) teilgenommen hatten. Der Retest wurde analog zum vorangegangenen Experiment mit Ausnahme von zwei Veränderungen durchgeführt: Als weiterer Sensor kam neben dem Emotiv EEG und dem Eyetracker Tobii X120 der BioHarness hinzu. Ein weiterer

Unterschied bezieht sich auf die Variation des Faktors *Lärm*. In Experiment 1 hatte sich gezeigt, dass der verwendete Maschinenlärm die Leistung und den Nutzerzustand nicht signifikant beeinflusst hat. Da sich die Untersuchungszeit bei zusätzlicher Variation der Lärmbedingung verdoppelt hätte, wurde aus Rücksicht auf die Teilnehmer auf eine Variation dieses Faktors im Retest verzichtet. Da manche Versuchsbedingungen und -teilnehmer wegen schlechter Aufzeichnungsqualität im ersten Experiment aus der Analyse ausgeschlossen werden mussten (vgl. Abschnitt 4.2.9), wurde jeweils die Versuchssitzung mit bzw. ohne Lärm im Retest wiederholt, für die im ersten Experiment eine bessere Aufzeichnungsqualität erzielt werden konnte. Dies führte dazu, dass 8 Teilnehmer den Retest ohne Lärm und 2 Teilnehmer den Retest mit Lärm durchführten.

Das experimentelle Design besteht im Retest somit aus den Faktoren *Anzahl Areas* und *Kooperativität der Kontakte*. Analog zum ersten Experiment wurden die Faktorstufen vollständig kombiniert, so dass vier Versuchsbedingungen resultieren (vgl. Abbildung 30). Um zu vermeiden, dass Reihenfolgeeffekte die Ergebnisse zur zeitlichen Stabilität beeinflussen, wurden die Versuchsbedingungen jedem Teilnehmer in der gleichen Reihenfolge präsentiert wie im ersten Experiment (vgl. Versuchsplan in Anhang A.1).

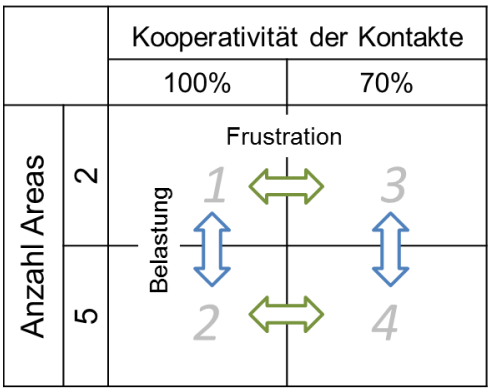


Abbildung 30. Experimentelles Design für die Retest-Untersuchung (Experiment 2)

5.2.2 Sensor Zephyr BioHarness 3

Der BioHarness ist ein physiologischer Multisensor, der zunächst für den athletischen Bereich entwickelt wurde und insbesondere auch im militärischen Kontext Anwendung findet. An einem Brustgurt befestigt (siehe Abbildung 31 links), erfasst der Sensor neben kardiovaskulären Maßen, wie *Herzrate* und *HRV* (berechnet wird der Parameter SDNN über ein 300-Sekunden-Intervall), auch die *Atemfrequenz, Temperaturveränderungen*, sowie *Körperhaltung* (Oberkörperbeugung relativ zur vertikalen Körperachse) und *Aktivität* über integrierte Beschleunigungssensoren. Die Daten werden drahtlos über Bluetooth auf einen Rechner übertragen und mittels eines SDK im CSV-Format gespeichert.

Verfügbare Maße des BioHarness

- Herzrate (Heartrate)
- Herzratenvariabilität (HRV)
- Atemfrequenz (Respiration)
- Körpertemperatur (Temperature)
- Körperhaltung (Posture)
- Aktivitätsniveau (Activity)

Abbildung 31. Multisensor Zephyr BioHarness 3 mit einer Übersicht über die wesentlichen von dem Sensor erfassten physiologischen und verhaltensbasierten Maße (Originalbezeichnung in Klammern)

5.2.3 Hypothesen

Hypothesen zur zeitlichen Stabilität.

Die Hypothesen zur Testung der ersten Forschungsfrage zur zeitlichen Stabilität der Eyetracking- und EEG-Maße sind in Tabelle 34 zusammengefasst. In Bezug auf Hypothese H1 wird angenommen, dass die Test-Retest-Korrelation (Korrelation der Messwertpaare aus erstem und zweitem Experiment pro Versuchsbedingung, siehe nähere Ausführungen dazu in Abschnitt 5.2.4) für jedes Diagnosemaß hoch positiv ausfällt. Lienert & Raatz (1998) schlagen für Testverfahren als Grenzwert für eine ausreichende Reliabilität eine Retestkorrelation von $r = .7$ vor, die nicht unterschritten werden sollte. Aus Literaturbefunden (z.B. Tomarken, 1995; Faulstich et al., 1986) geht jedoch hervor, dass die Test-Retest-Korrelationen bei physiologischen Maßen aufgrund der vielfältigen Einflussfaktoren, die sich auf die Messergebnisse in Test und Retest auswirken können, oftmals geringer ausfallen. Demzufolge kann eine positive Korrelation, die auf einen großen Zusammenhang hinweist ($r > .5$) bereits als zufriedenstellend erachtet werden (vgl. Llorente et al., 2001).

Tabelle 34. Hypothesen zur zeitlichen Stabilität für die Eyetracking- und EEG-Maße (Experiment 2)

Untersuchte Variablen	Untersuchte Zusammenhänge	Hypothese
- Pupillenweite - Fixationsdauer - EEG-Engagement - EEG-Frustration	Test-Retest-Korrelationen (Zusammenhänge zwischen den Messergebnissen einer Variable im erstem Experiment und im Retest)	H1: $\rho > .5$
- Pupillenweite - Fixationsdauer - EEG-Engagement - EEG-Frustration	Zusammenhänge zwischen den Diagnosemaßen und der Leistung im ersten Test (1) und im Retest (2)	H2a: $\rho_1 = \rho_2$
- Pupillenweite - Fixationsdauer - EEG-Engagement	Zusammenhänge zwischen den Diagnosemaßen und dem NASA-TLX-Gesamtscore im ersten Test (1) und im Retest (2)	H2b: $\rho_1 = \rho_2$
- EEG-Frustration	Zusammenhang zwischen EEG-Frustration und der Subskala Frustration des NASA-TLX im ersten Test (1) und im Retest (2)	H2c: $\rho_1 = \rho_2$

Von besonderem Interesse für die vorliegende Arbeit ist des Weiteren die Frage, ob Veränderungen in Nutzerzustand und Leistung reliabel durch die physiologischen Parameter erfasst werden können. In Experiment 1 zeigte sich, dass die Maße *Pupillenweite*, *Fixationsdauer* und *EEG-Engagement* statistisch signifikante Korrelationen mit der *subjektiven Beanspruchung* (NASA-TLX) und der *Leistung* (Punktestand) aufweisen. Um zu untersuchen, ob die gefundenen Zusammenhänge zeitlich stabil sind, wurden, ähnlich dem Vorgehen in „cross-lagged-panel"-Untersuchungen, die Eyetracking- und EEG-Maße mit dem Nutzerzustand und der Leistung zu zwei verschiedenen Zeitpunkten (Experiment 1 und Retest) an der gleichen Stichprobe korreliert (vgl. Bortz, 2005, S.223). Ausgehend davon, dass die Zusammenhänge zeitlich stabil sind, wird in den Hypothesen H2a, b und c in Tabelle 34 angenommen, dass keine Unterschiede in der Höhe der Korrelationen zwischen Experiment 1 und dem Retest bestehen. Hierbei ist die Nullhypothese auch die Wunschhypothese, was bei der Hypothesentestung zu berücksichtigen ist (siehe Abschnitt 5.2.4). In Hinblick auf die personenspezifische Eignung der Diagnosemaße wurde zudem deskriptiv untersucht, ob sich die im ersten Experiment identifizierten interindividuellen Unterschiede in der Höhe der Individualkorrelationen durch den Retest replizieren lassen.

Hypothesen zu den Maßen des BioHarness

Tabelle 35 gibt für die in Abbildung 31 aufgeführten Maße des BioHarness an, wie diese Literaturbefunden zufolge mit den Nutzerzuständen *Beanspruchung* und *Frustration* in Zusammenhang stehen (siehe hierzu auch Abschnitt 2.3.3). Auf Basis dieser Befunde wird angenommen, dass die Maße in Hinblick auf den Faktor Anzahl Areas signifikante Unterschiede aufweisen. Konkret wird erwartet, dass in Bedingungen mit 5 Areas und somit hoher Belastung die *Herzrate* und die *Atemfrequenz* höher ausfallen als in Bedingungen mit 2 Areas. Die *HRV* sollte hingegen bei 5 Areas niedriger sein als bei 2 Areas. Bei der *Körpertemperatur* stellten die in Tabelle 35 als Referenz angegebenen Studien ebenfalls ein Absinken der Temperatur bei Beanspruchung fest, wobei zu beachten ist, dass die Temperatur in diesen Studien nicht am Körper sondern an der Nase erfasst wurde. Für die *Körperhaltung* wird entsprechend der Literaturbefunde von Balaban et al. (2005) angenommen, dass der Oberkörper in Bedingungen mit 5 Areas stärker nach vorne geneigt ist als bei 2 Areas.

Darüber hinaus wird angenommen, dass die Maße Herzrate, HRV und Atemfrequenz auf einen durch Frustration ausgelösten höheren Erregungszustand reagieren. Entsprechend der Befunde sollten die *Herzrate* und die *Atemrequenz* in Bedingungen mit geringer Kooperativität höhere Werte aufweisen als in Bedingungen mit hoher Kooperativität. Für die *HRV* wird der umgekehrte Zusammenhang angenommen. Für das *Aktivitätsniveau* wird vermutet, dass sich dieses bei sitzender Tätigkeit nicht maßgeblich verändert. Dieses Maß wird daher nur zu exploratorischen Zwecken in die Analyse aufgenommen.

Außerdem wird angenommen, dass die Maße des BioHarness entsprechend den in Tabelle 35 formulierten Hypothesen mit den subjektiven Maßen zur Beanspruchung und Frustration korrelieren. Zusätzlich wird auch der Zusammenhang mit der Leistung untersucht. Es wird angenommen, dass hohe Beanspruchung und Frustration die Leistung verschlechtern. Die Maße sollten daher in entgegengesetzter Richtung mit der Leistung korrelieren als mit der Beanspruchung und der Frustration.

Tabelle 35. Hypothetisierte Zusammenhänge für die Maße des BioHarness (Experiment 2)

Maße des BioHarness	Korrelation mit Nutzerzustand		Referenz	Korrelation mit Leistung
Herzrate	Beanspruchung	$\rho > 0$	De Rivecourt, Kuperus & Mulder (2008); Vogt, Hagemann & Kastner (2006)	$\rho < 0$
	Frustration	$\rho > 0$		
Herzratenvariabilität (HRV)	Beanspruchung	$\rho < 0$	De Rivecourt, Kuperus & Mulder (2008); Backs & Seljos (1994)	$\rho > 0$
	Frustration	$\rho < 0$		
Atemfrequenz	Beanspruchung	$\rho > 0$	Backs & Seljos (1994); Backs & Ryan (1992); Wientjes (1992)	$\rho < 0$
	Frustration	$\rho > 0$		
Körpertemperatur	Beanspruchung	$\rho < 0$	Or & Duffy (2007); Veltman & Vos (2005) bzgl. Gesichtstemperatur (Nase)	$\rho > 0$
Oberkörperneigung	Beanspruchung	$\rho > 0$	Balaban et al. (2005)	$\rho < 0$
Aktivitätsniveau	Kein Zusammenhang im vorliegenden Kontext, da nur bei körperlicher Bewegung relevant.			

5.2.4 Datenaufbereitung und Auswertung

Die Daten wurden für den Retest in gleicher Form aufbereitet wie im ersten Experiment (vgl. Abschnitt 4.2.9). Bei den EEG-Maßen mussten die Daten von vier Personen wegen schlechter Aufzeichnungsqualität aus der Analyse ausgeschlossen werden. Somit beziehen sich die Ergebnisse für die EEG-Metriken auf n=6. Personen Bei den Eyetracking-Metriken betrifft dies zwei Personen, so dass die Analysen auf einer reduzierten Stichprobengröße von n=8 basieren.

Um die Hypothesen zur zeitlichen Stabilität zu untersuchen, wurden für die vier Eyetracking- und EEG-Maße, die im ersten Experiment untersucht wurden, Test-Retest-Korrelationen berechnet. Bei der Untersuchung auf Gruppenebene wurden die pro Versuchsbedingung vorliegenden Wertepaare von allen Personen in die Analyse einbezogen. Durch die z-Standardisierung der Variablen ist eine Bereinigung von Zwischensubjekteffekten durch Partialkorrelationen, wie in Abschnitt 4.2.10 beschrieben, nicht notwendig. Bei den Test-Retest-Korrelationen auf Individualebene wurden die pro Versuchsbedingung vorliegenden Wertepaare der Einzelpersonen herangezogen. Da der Retest nur vier Versuchsbedingungen enthält, liegt die Fallzahl bei n=4.

An dieser Stelle sei angemerkt, dass Bland & Altman (1986, 1996) zufolge Test-Retest-Korrelationen nicht ausreichen würden, um die Reliabilität bzw. zeitliche Stabilität eines Maßes zu bestätigen, da diese lediglich aussagen würden, ob ein Zusammenhang zwischen den Messungen bestehe, jedoch nicht, ob die Messungen äquivalent seien. Zum Beispiel könnten im Retest durchweg höhere Messergebnisse vorliegen, ohne dass dies den Zusammenhang reduziere. Alternativ besteht die Möglichkeit z.B. den Intraklassenkoeffizienten *ICC* (Shrout & Fleiss, 1979) zu berechnen, um die Äquivalenz und Reliabilität zu prüfen (Bland & Altman, 1996). Im Vordergrund der vorliegenden Untersuchung steht jedoch weniger die Frage, ob die absoluten Werte äquivalent sind (diese könnten zum Beispiel durchaus aufgrund der höheren Expertise im Retest abweichen), sondern ob Veränderungen in Abhängigkeit der Anforderungssituation reliabel

wiedergegeben werden. Aus diesem Grund wurden nicht die absoluten Werte sondern die z-standardisierten Werte in die Analyse einbezogen. Hierbei zeigte sich, dass der ICC genauso hoch wie die Produkt-Moment-Korrelation ausfällt. Bland & Altman (1996) geben außerdem zu Bedenken, dass die Reliabilität bei einer hohen Variabilität zwischen den Personen durch die Produkt-Moment-Korrelation überschätzt wird. Dies kann durch die vorgenommene Standardisierung ebenfalls vermieden werden.

Um zu überprüfen, ob die ermittelten Ergebnisse zum Zusammenhang der Eyetracking- und EEG-Maße mit der Leistung und der subjektiven Bewertung (NASA-TLX) zeitlich stabil sind, wurde getestet, ob die Korrelationen der Maße im Retest signifikant von den Korrelationen im ersten Experiment abweichen. Da es sich um Korrelationen handelt, die auf der gleichen Stichprobe und den gleichen Merkmalen basieren (sogenannte abhängige Korrelationen), wurde zur Berechnung der standardnormalverteilten Prüfgröße Z die von Raghunatan, Rosenthal, & Rubin (1996) modifizierte *Pearson-Filon-Statistik (ZPF)* angewendet. Im Unterschied zu der Formel von Pearson-Filon (1898) erfolgt die Berechnung auf Basis der in Fisher's Z-Werte transformierten Korrelationen. Zur Berechnung der Prüfgröße Z wurde das Softwaretool *SISA* von Uitenbroek (1997) genutzt. Da es sich bei der Nullhypothese ($\rho_1 = \rho_2$) um die zu überprüfende Wunsch-hypothese handelt, wurde – als Maßnahme um den β-Fehler zu reduzieren – der Empfehlung von Döring & Bortz (2016, S.885) folgend das alpha-Niveau auf $\alpha = .10$ hochgesetzt.

Die Maße des BioHarness wurden analog zu den Eyetracking- und EEG-Maßen in Experiment 1 z-standardisiert und varianzanalytisch in Bezug zu den Faktoren Anzahl Areas und Kooperativität ausgewertet. Zusammenhänge mit der Leistung und dem subjektiven Nutzerzustand wurden wie im ersten Experiment über Produkt-Moment-Korrelationen berechnet. Es mussten keine Daten wegen schlechter Aufzeichnungsqualität ausgeschlossen werden, so dass die Auswertungen auf einer Stichprobengröße von N=10 basieren.

5.3 Ergebnisse zur zeitlichen Stabilität der physiologischen und verhaltensbasierten Maße

5.3.1 Test-Retest Korrelationen

Tabelle 36 führt die Test-Retest-Korrelationen für die betrachteten Eyetracking und EEG-Maße auf Gruppenebene auf. Zudem sind in Abbildung 32 die Test-Retest-Korrelationen pro Person für die Eyetracking- und EEG-Maße als Balkendiagramme dargestellt. Der jeweils erste Balken zeigt zum besseren Vergleich mit dem Resultat auf Gruppenebene den Mittelwert der Individualkorrelationen an. Zur Berechnung des Mittelwerts wurden die Korrelationen zuvor in Fisher's Z-Werte transformiert und der Mittelwert wurde anschließend in r rücktransformiert (vgl. Bortz, 2005, S. 219).

Auf Gruppenebene zeigen sich für die *Pupillenweite*, die *Fixationsdauer* und *EEG-Frustration* signifikante Test-Retest-Korrelationen, die mit $r > .5$ kongruent zu Hypothese H1 ausfallen. *EEG-Engagement* weist hingegen mit $r = .36$ eine deutlich geringere Test-Retest-Korrelation auf; H1 lässt sich somit für dieses Maß nicht bestätigen. Im Vergleich fallen die Mittelwerte der Individual-korrelationen höher aus als die Korrelationen auf Gruppenebene. In den meisten Fällen weisen die Korrelationen mit $r > .5$ auf einen starken Zusammenhang hin. Dies spricht für eine weitgehende

Stabilität der Maße auf individueller Ebene. Dennoch ist zu beachten, dass bei manchen Personen, insbesondere bei *EEG-Engagement*, auch keine oder negative Test-Retest-Korrelationen vorliegen. Außerdem ist festzustellen, dass die Höhe der Individualkorrelationen nicht nur zwischen den Teilnehmern sondern auch pro Teilnehmer zwischen den Maßen teilweise stark schwankt. Zum Beispiel weist VP 9 eine hohe positive Test-Retest-Korrelation für die Fixationsdauer aber eine nicht erwartungskonforme negative Korrelation für die Pupillenweite auf.

Tabelle 36. Test-Retest Korrelationen der Eyetracking- und EEG-Maße auf Gruppenebene (Experiment 2)

	Pupillenweite	Fixationsdauer	EEG-Engagement	EEG-Frustration
Korrelation *r*	.65**	.53**	.36	.63**
Fallzahl	36	36	24	24

* $p<.05$, ** $p<.01$ bei einseitigem Testen.

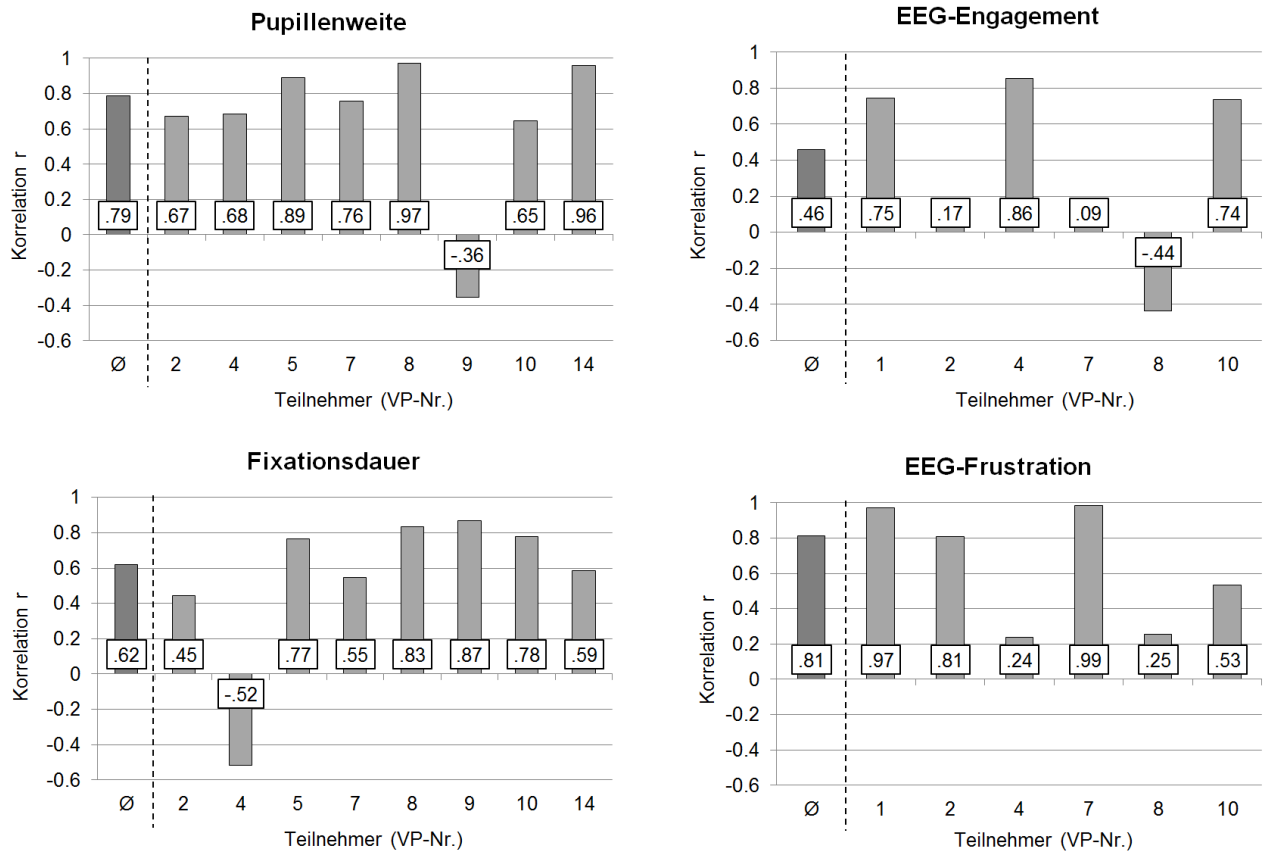


Abbildung 32. Test-Retest-Korrelationen für die Eyetracking- und EEG-Maße pro Teilnehmer (Ø = Mittelwert der Individualkorrelationen) – Experiment 2

5.3.2 Korrelation der Eyetracking- und EEG-Maße mit der Leistung und dem subjektiven Nutzerzustand in Test und Retest

Tabelle 37 führt die Zusammenhänge der Eyetracking- und EEG-Maße mit dem Punktestand als Leistungsmaß sowie dem NASA-TLX gesondert für das erste Experiment und den Retest auf Gruppenebene auf. Abbildung 33 veranschaulicht die Korrelationen der Eyetracking- und EEG-Maße mit dem Punktestand auf Individualebene als Balkendiagramme. Neben den Individualkorrelationen sind links jeweils die über Fisher's Z-Transformation ermittelten Mittelwerte der Individualkorrelationen dargestellt (rot markiert).

Tabelle 37. Statistische Kennwerte zu den Korrelationen der Eyetracking- und EEG-Maße mit der Leistung (Punktestand) und dem NASA-TLX auf Gruppenebene im ersten Experiment und im Retest (Experiment 2)

		Pupillenweite (Fallzahl *N*=32)	**Fixationsdauer** (Fallzahl *N*=32)	**EEG-Engagement** (Fallzahl *N*=24)	**EEG-Frustration** (Fallzahl *N*=24)
Korrelation mit Punktestand	1. Experiment / Retest	-.34*/-.36*	.60**/.37*	-.57**/ .21	-.26/-.40*
	Differenz	.02	.23	.78**	.14
	ZPF	0.15	1.58	-3.50	0.75
Korrelation mit NASA-TLX[a]	1. Experiment / Retest	.60**/.42*	-.61**/-.42*	.65**/ -.23	.23/.48*
	Differenz	.18	.19	.88**	.25
	ZPF	1.29	-1.28	4.18	1.21

* $p<.05$, ** $p<.01$; [a] bezieht sich auf den NASA-TLX-Gesamtscore, bzw. für EEG-Frustration auf die Subskala Frustration. ZPF: Prüfgröße Z, ermittelt über Pearson-Filon-Verfahren nach Raghunathan et al. (1996)

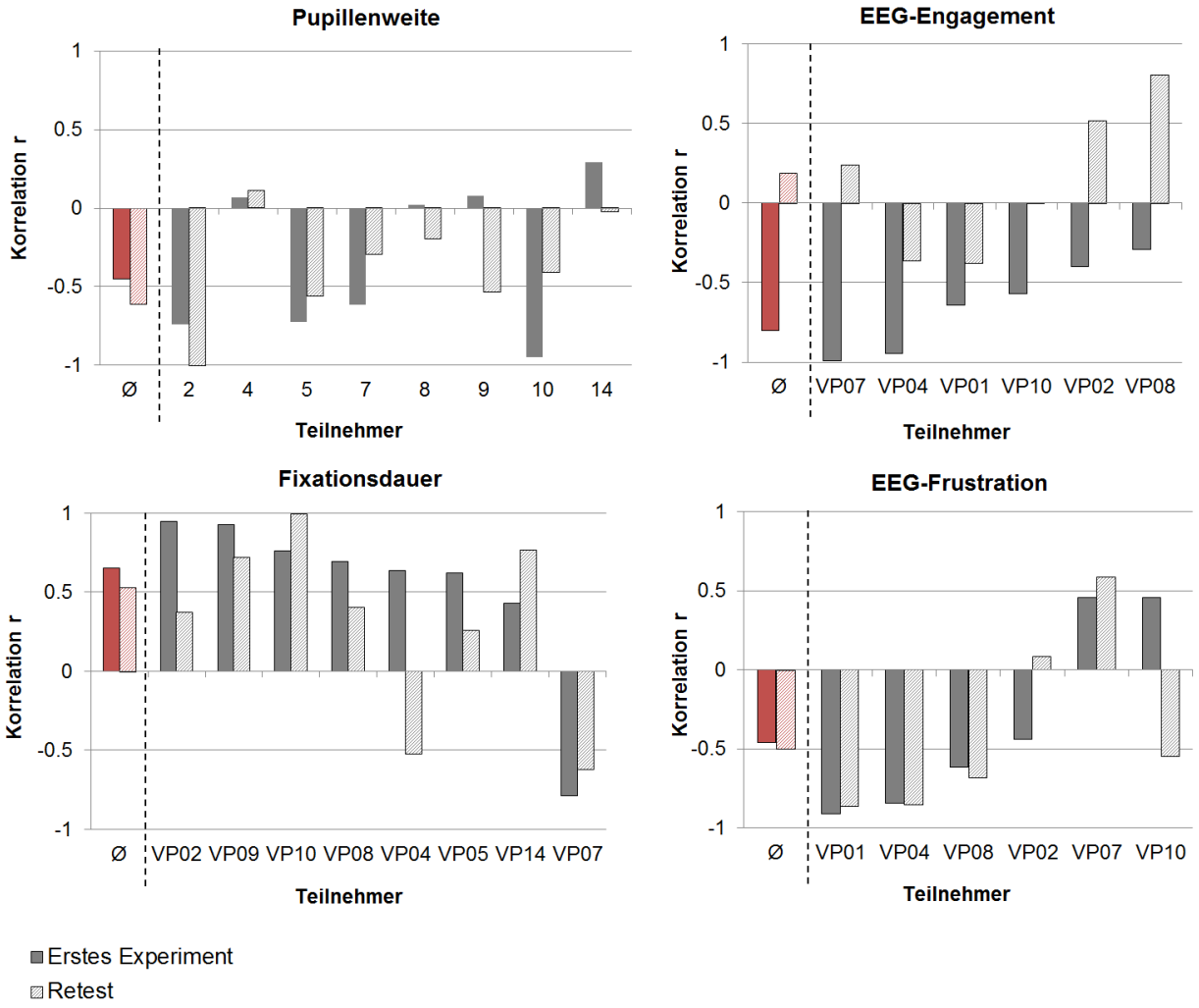


Abbildung 33. Individualkorrelationen und Mittelwert (Ø) der Eyetracking- und EEG-Maße mit der Leistung (Punktestand) für das erste Experiment[6] und den Retest (Experiment 2)

[6] Da für den Test-Retest-Vergleich die Daten nur einer Versuchssitzung aus Experiment 1 herangezogen wurden, sind die Individualkorrelationen für Experiment 1 nicht identisch mit den Korrelationen in Abbildung 28.

Auf Gruppenebene ist festzustellen, dass die Maße *Pupillenweite* und *Fixationsdauer* in beiden Experimenten signifikante Korrelationen mit der Leistung (Punktestand) und dem NASA-TLX aufweisen, die jeweils in erwarteter Richtung ausfallen. Die Unterschiede zwischen den Korrelationen erweisen sich für diese Maße auf dem 10%-Niveau als nicht signifikant. Die Hypothesen 2a und b können für diese Maße somit auf Gruppenebene bestätigt werden.

In Hinblick auf *EEG-Engagement* unterscheiden sich die Korrelationen mit dem Punktestand und dem NASA-TLX im ersten Experiment und im Retest hingegen signifikant ($p<.01$). Während EEG-Engagement im ersten Experiment erwartungsgemäß negativ mit dem Punktestand und positiv mit dem NASA-TLX korreliert, liegen im Retest schwache Korrelationen in umgekehrter Richtung vor. Die Hypothesen 2a und b können für EEG-Engagement daher nicht bestätigt werden. Bei *EEG-Frustration* fallen die Korrelationen im ersten Experiment deutlich schwächer aus als im Retest. Diese Unterschiede erweisen sich jedoch auf dem 10%-Niveau als nicht signifikant, so dass die Hypothesen 2a und c für dieses Maß als bestätigt angenommen werden können.

Die Mittelwerte der Individualkorrelationen sind in beiden Experimenten etwas höher als die Korrelationen auf Gruppenebene und unterscheiden sich für die *Pupillenweite*, die *Fixationsdauer* und *EEG-Frustration* zwischen dem ersten Experiment und dem Retest nur geringfügig. Bei *EEG-Engagement* liegen hingegen, wie der Befund auf Gruppenebene bereits anzeigte, starke Unterschiede zwischen dem ersten Test und dem Retest vor.

Insgesamt wird ersichtlich, dass die Höhe der Individualkorrelationen bei den betrachteten Maßen nicht nur zwischen den Teilnehmern stark schwankt sondern sich auch innerhalb der Teilnehmer vom ersten Experiment zum Retest unterscheidet. Beispielsweise weist die Fixationsdauer bei VP 2 im ersten Experiment eine sehr hohe Korrelation mit der Leistung auf. Im Retest fällt die Korrelation hingegen vergleichsweise gering aus. Dies deutet darauf hin, dass der Zusammenhang mit der Leistung nicht nur interindividuell unterschiedlich ist, sondern sich auch zwischen den Erhebungszeitpunkten unterscheiden kann. Die zeitliche Stabilität kann somit auf Individualebene nicht bestätigt werden.

5.4 Ergebnisse zu den Maßen des BioHarness

Die Ergebnisse der varianzanalytischen Untersuchung zum Verhalten der Maße des BioHarness in Abhängigkeit der experimentell modulierten Anforderungsmerkmale Anzahl Areas und Kooperativität der Kontakte sind in Abbildung 34 dargestellt. Signifikante Haupteffekte ergeben sich nur für die *HRV* ($F(9,1)=8.552$, $p<.05$) und die *Atemfrequenz* ($F(9,1)=6.834$, $p<.05$) in Hinblick auf den Faktor *Anzahl Areas*. Erwartungskonform ist die HRV in der Bedingung mit 5 Areas (hoher Belastung) niedriger als in der Bedingung mit 2 Areas (geringer Belastung). Die Atemfrequenz fällt – ebenfalls erwartungskonform – in der Bedingung mit hoher Belastung höher aus als in der Bedingung mit geringer Belastung. Für die übrigen Maße können zumeist nur sehr schwache Unterschiede zwischen den Faktorstufen festgestellt werden, die sich als nicht signifikant und nicht praktisch bedeutsam erweisen.

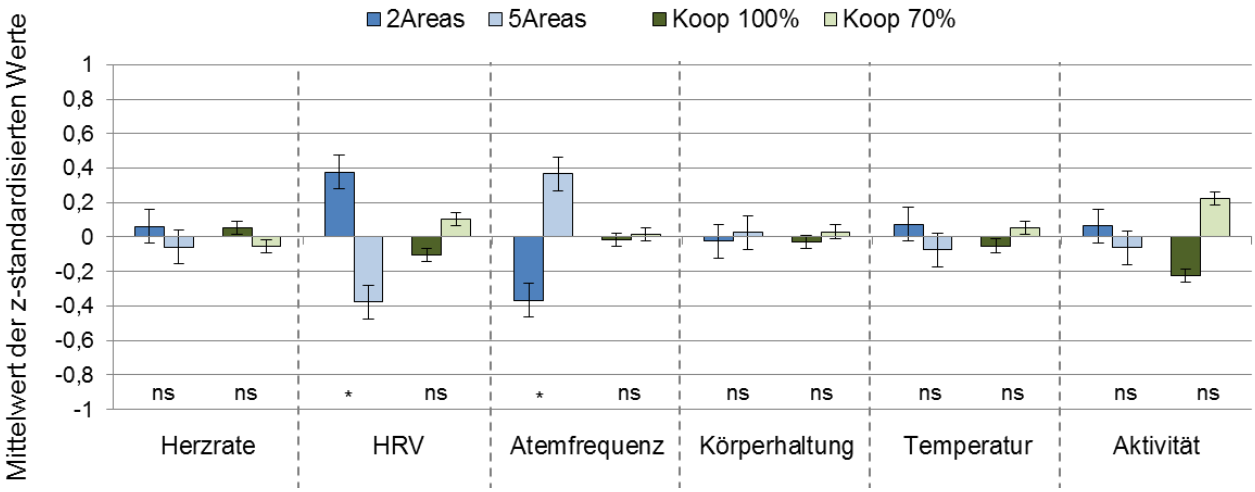


Abbildung 34. Mittelwerte und Standardfehler pro Faktorstufe für die Maße des BioHarness in Experiment 2 (Fehlerbalken: Standardfehler; *p<.05; ns – nicht signifikant)

Tabelle 38 gibt die Zusammenhänge der Maße des BioHarness mit der subjektiven Beanspruchung und der Frustration (jeweils NASA-TLX) sowie mit der Leistung (Punktestand) wieder. Auch hier zeigt sich, dass die *HRV* und die *Atemfrequenz* starke signifikante Zusammenhänge in erwarteter Richtung mit den subjektiven Maßen und der Leistung aufweisen. Die Zusammenhänge der übrigen Maße fallen zumeist sehr schwach aus. Eine Ausnahme bildet die *Körpertemperatur*. Sie korreliert signifikant negativ mit der Subskala Frustration und weist einen moderaten negativen Zusammenhang mit dem NASA-TLX-Gesamtscore auf.

Tabelle 38. Korrelationen der Maße des BioHarness mit dem NASA-TLX und dem Punktestand (Leistung) auf Gruppenebene nach Bereinigung von Zwischensubjekteffekten (Experiment 2)

	Herzrate	HRV	Atemfrequenz	Körperhaltung	Temperatur	Aktivität
Korrelation mit Punktestand	.01	.49**	-.37*	.05	.17	.06
Korrelation mit NASA-TLX	-.1	-.50**	.39*	-.05	-.32	-.09
Korrelation mit Frustration[a]	-.1	-.39*	.41*	.01	-.44**	-.06

* p<.05, **p<.01 bei einseitigem Testen; [a] bezieht sich auf die NASA-TLX Subskala Frustration, Fallzahl N=40

5.5 Diskussion

In den Analysen zur zeitlichen Stabilität der Eyetracking- und EEG-Maße zeigte sich, dass diese mit Ausnahme von EEG-Engagement auf Gruppenebene eine hohe Test-Retest-Stabilität aufweisen. Die Test-Retest-Korrelationen liegen trotz des hohen Zeitabstandes zwischen den Erhebungen auf einem ähnlichen Niveau wie sie in den betrachteten Studien von Tomarken (1995) sowie in der Studie von Faulstich et al. (1986) bei Verwendung der absoluten Werte ermittelt wurden. Die Analyse der Zusammenhänge der Eyetracking- und EEG-Maße mit dem NASA-TLX und dem Punktestand bestätigte ebenfalls eine weitgehende Stabilität der Ergebnisse für die Pupillenweite, die Fixationsdauer und EEG-Frustration auf Gruppenebene.

Für *EEG-Frustration* ist dieses Ergebnis bemerkenswert, da dieses Maß im ersten Experiment keine bedeutsamen Korrelationen mit den subjektiven Maßen und den Leistungsmaßen aufwies (vgl. Abschnitt 4.3.3). Hierbei ist zu beachten, dass für den Vergleich mit dem Retest nur die Teilnehmer und Versuchsbedingungen in die Analyse eingeschlossen werden konnten, für die auch im Retest Daten vorliegen. Die Reduzierung der Fallzahl führte offensichtlich dazu, dass weniger valide Messungen ausgeschlossen wurden und EEG-Frustration nun auch für das erste Experiment zumindest mäßig hohe Korrelationen in erwarteter Richtung mit der subjektiv bewerteten Frustration und der Leistung aufweist.

Ein umgekehrtes Ergebnis liegt für das *EEG-Engagement* vor. Die Ergebnisse aus dem ersten Experiment bestätigten die Validität dieses Maßes für die Erfassung der mentalen Beanspruchung. Beim Vergleich der Korrelationen mit dem NASA-TLX und dem Punktestand zwischen dem ersten Experiment und dem Retest fällt jedoch auf, dass für EEG-Engagement im Retest nur schwache und nicht erwartungskonforme Korrelationen vorliegen. Da dieser Befund bei den anderen Maßen nicht zu beobachten ist, ist zu vermuten, dass die Aufzeichnung des EEG-Engagements im Retest durch Störeinflüsse beeinträchtigt wurde. Leider bietet die Klassifikationssoftware des Emotiv EEG keine Informationen zu der Qualität der Messwertaufzeichnung. Die Identifikation nicht valider Daten und der Ausschluss dieser Daten aus der Analyse wird dadurch erschwert (vgl. hierzu Abschnitt 4.2.9).

Auf Individualebene konnten für die meisten Teilnehmer ebenfalls mittlere bis hohe positive Test-Retest-Korrelationen ermittelt werden. Dennoch ist zu beachten, dass in einigen Fällen auch niedrige oder negative Test-Retest-Korrelationen vorliegen. Bei den Korrelationen der Eyetracking- und EEG-Maße mit dem NASA-TLX und dem Punktestand zeigte sich, dass sich diese bei vielen Teilnehmern zwischen dem ersten Experiment und dem Retest stark unterscheiden. Vereinzelt unterscheidet sich nicht nur die Höhe sondern auch die Richtung der Korrelationen in beiden Experimenten. Bei der Interpretation dieser nicht erwartungskonformen individuellen Ergebnisse ist zu bedenken, dass durch die geringe Fallzahl pro Person von n=4 ein einzelner nicht valider Wert bereits erhebliche Auswirkungen auf die Höhe und Richtung der Korrelation hat. Um die Zusammenhänge auf individueller Ebene genauer zu untersuchen und gegen den Zufall abzusichern, wäre eine größere Fallzahl an Beobachtungen pro Individuum erforderlich.

Trotz der fallzahlbedingt eingeschränkten Aussagekraft weisen die heterogenen Ergebnisse insgesamt aber bereits darauf hin, dass es nicht nur vom Individuum sondern auch vom Erhebungszeitpunkt abhängt, wie gut ein Maß auf individueller Ebene Unterschiede im Nutzerzustand und der Leistung anzeigen kann. Tomarken (1995) führt als Erklärung für beobachtete Unterschiede in der zeitlichen Stabilität zwischen physiologischen Maßen an, dass die Erfassung durch verschiedene Störvariablen, wie die Verfassung der Versuchsteilnehmer und die Konsistenz in der Platzierung von Elektroden beeinflusst wird. Diese Faktoren haben vermutlich auch in der vorliegenden Untersuchung dazu beigetragen, dass sich die Ergebnisse nicht nur zwischen Personen sondern auch bei der gleichen Person von einem Erhebungszeitpunkt zum nächsten unterscheiden.

In Bezug auf die Maße des BioHarness zeigte sich, dass sich die *HRV* und die *Atemfrequenz* signifikant zwischen hoher und niedriger Belastung unterscheiden und daher als Indikatoren für die

mentale Beanspruchung in Betracht kommen. In Hinblick auf den Faktor Kooperativität weisen die Maße des BioHarness – wie auch die übrigen physiologischen Maße – keine signifikanten Unterschiede auf. Daher ist zu vermuten, dass die Bedingung mit geringer Kooperativität den über physiologische Maße erfassbaren Nutzerzustand nicht stark genug beeinflusst hat. Dennoch zeigte sich, dass die *HRV*, die *Atemfrequenz* und die *Körpertemperatur* signifikant mit der subjektiv erfassten Frustration korrelieren. Zu berücksichtigen ist allerdings, dass diese Maße geringe Diagnostizität besitzen, da sie ebenfalls mit der mentalen Beanspruchung korrelieren. Für eine Echtzeitdiagnose in Realumgebungen sind daher weitere Informationen notwendig, um zwischen den verschiedenen mentalen Zuständen diskriminieren zu können. Für verallgemeinerbare Aussagen zur Stärke der Zusammenhänge ist außerdem eine Untersuchung mit einer größeren Stichprobe notwendig.

5.6 Resümee

Insgesamt sprechen die Ergebnisse des Retests dafür, dass die Eyetracking-Maße *Pupillenweite* und *Fixationsdauer* sowie die Maße *HRV* und *Atemfrequenz* des BioHarness einen Beitrag leisten können, den Nutzerzustand im Rahmen einer multifaktoriellen Diagnose zu erfassen. Eine reliable und valide Erfassung des Nutzerzustands erscheint mit dem Emotiv-EEG allerdings nicht möglich. Der Klassifikator *EEG-Frustration* hat sich bereits in Experiment 1 als nicht valide herausgestellt. Die Ergebnisse im Retest stellten außerdem die Validität und zeitliche Stabilität des *Engagement*-Klassifikators in Frage. Darüber hinaus mussten in beiden Experimenten mehrere Teilnehmer aufgrund ungenügender Aufzeichnungsqualität aus der Analyse ausgeschlossen werden.

Die individuellen Unterschiede in der zeitlichen Stabilität verdeutlichen außerdem, dass Ergebnisse auf Gruppenebene, wie sie üblicherweise in der psychologischen Forschung berichtet werden, nur bedingt aussagekräftig sind, um zu entscheiden, ob ein Diagnoseinstrument im Einzelfall geeignet ist, den Nutzerzustand valide und verlässlich zu erfassen. Veltman & Jansen (2003) empfehlen daher eine personenspezifische Auswahl von Diagnosemaßen (vgl. Abschnitt 2.4.3). Allerdings stellte sich im Retest auch heraus, dass die Sensitivität eines Diagnosemaßes nicht nur interindividuell sondern auch intraindividuell zwischen den Erhebungszeitpunkten unterschiedlich ausfallen kann. Diese unsystematischen zeitpunktabhängigen Schwankungen zeigten sich im Retest für die untersuchten vier physiologischen Maße und sind den Ergebnissen von Uhlig (2018) zufolge auch für die HRV zu erwarten.

Eine nicht nur nutzer-, sondern auch zeitpunktspezifische Bestimmung geeigneter Diagnosemaße ist für die Erfassung des Nutzerzustands in realen Anwendungsfällen allerdings nicht praktikabel. Insofern erscheint es angebracht, Ungenauigkeiten bei einzelnen Maßen durch Kombination verschiedener Maße auszugleichen und auf diese Weise ein robusteres Diagnoseergebnis zu erzielen (vgl. Abschnitt 2.4.2).

6 Konzeption und Umsetzung der multifaktoriellen Echtzeitdiagnose RASMUS

Auf Basis der Literaturbefunde (vgl. Kapitel 2 und 3) und den Erkenntnissen aus den in Kapitel 4 und 5 beschriebenen experimentellen Untersuchungen wurde ein generisches Konzept für eine multifaktorielle Echtzeitdiagnose des Nutzerzustands entwickelt. Diese trägt den Kurznamen *RASMUS*, ein Akronym für *Real-time Assessment of Multidimensional User State*. In diesem Kapitel wird das Diagnosekonzept zunächst in allgemeiner Form vorgestellt (Abschnitt 6.1). Daran schließt eine detaillierte Beschreibung des Diagnoseprozesses in Abschnitt 6.2 an. Die darauffolgenden Abschnitte beziehen sich auf die Umsetzung der Echtzeitdiagnose für einen Anwendungsfall im Bereich der Luftraumüberwachung. Abschnitt 6.3 stellt das zugrunde liegende Experimentalparadigma und Abschnitt 6.4 die dabei betrachteten Nutzerzustände vor. Um Indikatoren und Regeln der Echtzeitdiagnose für den betrachteten Anwendungsfall zu identifizieren, werden die Experimentaldaten eines vorangegangenen Experiments herangezogen (vgl. Experiment 3 in Abschnitt 6.5). Die daraus abgeleiteten Indikatoren und Regeln werden in Abschnitt 6.6 erläutert. Abschnitt 6.7 geht schließlich auf die technische Umsetzung von RASMUS ein.

6.1 Diagnosekonzept

Übergeordnetes Ziel der Echtzeitdiagnose ist es, Informationen über Nutzer- und Umweltzustände in Echtzeit zu erfassen und zu bewerten, um so eine dynamische, bedarfsgerechte Adaptierung der Mensch-Maschine-Interaktion zu ermöglichen. Durch RASMUS soll es insbesondere möglich sein zu bestimmen, (1) wann eine Adaptierung notwendig ist, und (2) welche möglichen Ursachen für Leistungsminderungen und Problemzustände in Betracht kommen (vgl. Abschnitt 1.4). Diese Informationen werden anschließend von dem Adaptierungsmanagement (ADAM; vgl. Fuchs & Schwarz, 2017) weiter verarbeitet, um Auswahl und Konfiguration geeigneter Adaptierungsstrategien zu bestimmen. Im Folgenden wird vorgestellt, wie das Diagnosekonzept diese beiden zentralen Aufgaben adressiert.

6.1.1 Bestimmung von Adaptierungsbedarf

In der Literatur werden zumeist physiologische Maße als Auslöser für Adaptierungen herangezogen (vgl. Abschnitt 2.2). Wie nachfolgend näher ausgeführt wird, soll der Adaptierungsbedarf in RASMUS dagegen auf Basis von Leistungsveränderungen bestimmt werden. Eine Adaptierung wird somit erst dann ausgelöst, wenn es bereits zu einer Verschlechterung der Leistung – im Folgenden „Leistungseinbruch“ genannt – gekommen ist. Die Möglichkeit, Leistungseinbrüchen bereits vor ihrem Auftreten proaktiv entgegenzuwirken, ist bei dieser Vorgehensweise zwar nicht gegeben, jedoch sprechen zwei wesentliche Argumente für diesen Ansatz:

Zum Einen stellte sich in den vorangegangenen Kapiteln heraus, dass die Verwendung physiologischer Maße mit verschiedenen Herausforderungen verbunden ist. So zeigte sich, dass Störeinflüsse die Reliabilität von einzelnen Maßen beeinflussen können. Auch interindividuelle

Unterschiede sowie starke kurzzeitige Fluktuationen in den physiologischen Reaktionen können, wie in Abschnitt 2.4 dargelegt wurde, Probleme für eine verlässliche Erfassung des Nutzerzustands darstellen. Belege für diese theoriebasierten Erkenntnisse lieferten die Experimente 1 und 2 (vgl. Kapitel 4 und 5), in denen sich zeigte, dass die Sensitivität der physiologischen Indikatoren sowohl zwischen Personen, als auch innerhalb einer Person zu verschiedenen Zeitpunkten variierte. Adaptierungen, die dann auslösen, wenn ein Indikator einen vordefinierten Grenzwert unter- oder überschreitet, könnten daher unnötig oder inadäquat sein und beim Nutzer Irritationen hervorrufen. Dies gilt insbesondere dann, wenn Adaptierungen durch Oszillieren der Werte in kurzen Zeitabständen ausgelöst und wieder gestoppt werden (Scallen, Hancock, & Duley, 1995; Barker et al. 2004; Inagaki, 2003; vgl. Abschnitt 2.4.7). Das Vorliegen eines Leistungseinbruchs ist hingegen ein sehr sicheres Indiz, dass der Nutzer Unterstützung benötigt, so dass die Gefahr inadäquater Adaptierungen aufgrund von falsch positiven Diagnosen vermindert wird.

Zum Anderen spricht für dieses Vorgehen auch die Erkenntnis, dass der menschliche Organismus selbst ein adaptives System ist, das sich durch die Anwendung von Selbstregulierungsstrategien an Veränderungen der Anforderungssituation anpassen kann (vgl. Anforderung 4 in Tabelle 7, S. 33 und Abschnitt 2.4.4). Physiologische Reaktionen, die z.B. bei erhöhter Anstrengung auftreten, sind ein Zeichen des adaptiven Verhaltens. Eine Adaptierung auf Basis physiologischer Reaktionen könnte somit dazu führen, dass zwei adaptive Systeme kontraproduktiv zusammenarbeiten. Veltman & Jansen (2006) empfehlen daher, dass ein adaptives System erst dann eingreifen sollte, wenn der menschliche Operateur nicht mehr adaptierungsfähig ist. Ein Anzeichen für eine unzureichende Selbstadaptierung des Operateurs ist das Auftreten eines Leistungseinbruchs. Dabei sei jedoch angemerkt, dass für jeden Anwendungsfall individuell festgelegt werden muss, wie ein Leistungseinbruch definiert wird. Insbesondere bei sicherheitskritischen Systemen ist sicherzustellen, dass die Adaptierung nicht erst dann erfolgt, wenn es bereits zu spät ist. In diesem Fall könnte es erforderlich sein, bereits eine moderate Leistungsverschlechterung als Leistungseinbruch zu bewerten.

6.1.2 Bestimmung der Problemzustände und Ursachen

Leistungsmaße geben, wie in Abschnitt 2.3.2 erläutert wurde, keinen Aufschluss darüber, welche Faktoren und kritischen Nutzerzustände zu dem Leistungseinbruch beigetragen haben. Da das adaptive System diese Informationen aber benötigt, um eine zustands- und situationsadäquate Auswahl von Adaptierungsstrategien zu ermöglichen, sieht das Konzept von RASMUS vor, in einem weiteren Schritt Ursachen für das Auftreten von Leistungseinbrüchen zu bestimmen. Dazu analysiert RASMUS, welche Nutzerzustände potenziell kritisch ausgeprägt sind, und welche Einflussfaktoren daran beteiligt sind.

Als kritische Nutzerzustände oder Problemzustände gelten Ausprägungen der sechs in Kapitel 3 beschriebenen Nutzerzustandsdimensionen, welche sich negativ auf die menschliche Leistung auswirken. In den Kapiteln 4 und 5 wurden exemplarisch die Problemzustände *hohe mentale Beanspruchung*, *Frustration* und *eingeschränkte Aufmerksamkeit* untersucht. Daneben können auch andere negative Emotionen, wie Angst oder Stress, hohe Müdigkeit, geringe Motivation oder inadäquates Situationsbewusstsein die Leistung beeinträchtigen. Auch wenn es in der praktischen

Umsetzung nicht immer möglich ist, alle sechs Nutzerzustandsdimensionen zu betrachten, sollen diese im Rahmen der angestrebten ganzheitlichen Betrachtungsweise zumindest konzeptionell berücksichtigt werden. Für die praktische Umsetzung ist hingegen denkbar, nur die Zustände in die Diagnose einzubeziehen, die für das jeweilige Anwendungsfeld die höchste Relevanz aufweisen.

Als Grundlage für die Erfassung und Bewertung des Nutzerzustands dient das in Abschnitt 3.4 beschriebene und in Abbildung 15, S. 65 dargestellte generische Modell zum Nutzerzustand. Physiologische und verhaltensbasierte Maße werden durch den Nutzerzustand beeinflusst und bieten, wie bereits in Abschnitt 2.3.2 erläutert wurde, den Vorteil, dass sie kontinuierlich erfasst und in Echtzeit ausgewertet werden können. Zu berücksichtigen ist jedoch, dass diese Maße auf unterschiedliche Nutzerzustände und unterschiedliche Einflussfaktoren reagieren (vgl. Abschnitte 2.3.3 und 2.4.1). Dies zeigte sich in den Experimentalergebnissen u.a. daran, dass einige Maße (z.B. HRV und Atemfrequenz) sowohl mit der mentalen Beanspruchung als auch mit Frustration korrelieren. Außerdem lassen sie keine Aussagen zu möglichen Ursachen für kritische Nutzerzustände zu. Daher sieht das Diagnosekonzept vor, neben physiologischen und verhaltensbasierten Maßen, die im Modell (Abbildung 15) enthaltenen umweltbezogenen und individuellen Einflussfaktoren in die Bewertung des Nutzerzustands einzubeziehen. Die Kombination der verschiedenen Maße soll die Robustheit der Diagnose erhöhen und mögliche Ursachen für Leistungseinbrüche und kritische Nutzerzustände bereitstellen.

6.1.3 Überblick über das Diagnosekonzept

Die Nutzerzustandsbewertung in RASMUS erfolgt sequentiell und beginnt mit der Untersuchung von Leistungseinbrüchen. Als Indikatoren für einen Leistungseinbruch können z.B. die Dauer der Aufgabenbearbeitung oder das Auftreten von Fehlern herangezogen werden. Sofern ein Leistungseinbruch identifiziert wurde, folgt eine Analyse, welche Indikatoren des Nutzerzustands kritische bzw. ungünstige Ausprägungen aufweisen. Auf dieser Basis bestimmt RASMUS, welche Nutzerzustände als Ursachen oder Auslöser für den Leistungseinbruch in Betracht kommen (vgl. Abbildung 35).

Abbildung 35. Vorgehen bei der multifaktoriellen Nutzerzustandsdiagnose in RASMUS

Diese Analysen erfolgen im Rahmen der Promotionsarbeit regelbasiert. Alternativ besteht die Möglichkeit hierzu Maschinelles Lernen bzw. Methoden der Künstlichen Intelligenz einzusetzen (z.B. Artificial Neural Networks, vgl. Abschnitt 2.2.1). Das technische System ermittelt dabei selbst, durch welche Indikatoren und Kombinationen von Indikatorausprägungen kritische Nutzerzustände erfolgreich detektiert werden können. Voraussetzung ist die Verfügbarkeit großer Datenmengen, um die Algorithmen zu trainieren. Ein Nachteil besteht darin, dass es sich bei diesen

Systemen um so genannte „black boxes“ handelt, deren Verhalten zwar beobachtet, aber schwer nachvollzogen und begründet werden kann (Matthias, 2004). In der Promotionsarbeit wird hingegen angestrebt, den Diagnoseprozess transparent zu halten, so dass die den Diagnoseentscheidungen zugrunde liegenden Wirkzusammenhänge nachvollzogen werden können. Daher wird der Ansatz verfolgt, geeignete Indikatoren und Regeln zur Nutzerzustandsdiagnose aus den theoretischen und empirischen Analysen abzuleiten.

6.2 Beschreibung des Diagnoseprozesses

In Kapitel 1 wurde das selbst entwickelte generische Modell zur adaptiven MMI vorgestellt, das die grundsätzliche Funktionsweise des zu entwickelnden adaptiven technischen Systems beschreibt (Abbildung 1). Die zentrale Komponente, die das technische System zu adaptivem Verhalten befähigen soll, wird in diesem Modell als Zustandsregulierung bezeichnet. Das Diagnoseverfahren RASMUS bezieht sich auf die ersten beiden Stufen dieses Zustandsregulierungsprozesses: *Zustandsaufnahme* und *Zustandsbewertung*. Das Modell in Abbildung 36 führt diese beiden Stufen der Zustandsregulierung genauer aus. Die schwarzen Pfeile zeigen den Ablauf des Diagnoseprozesses an. Die grauen Pfeile weisen darauf hin, dass für die Durchführung eines Diagnoseschritts auch Informationen aus vorhergehenden Schritten benötigt werden.

Basierend auf dem Diagnosekonzept (vgl. Abschnitt 6.1) untergliedert sich der Diagnoseprozess in vier Schritte, die in Abbildung 36 als weiße Kästen dargestellt sind. Der erste Schritt bezieht sich auf die Stufe der Zustandserfassung und beinhaltet die Erfassung und Zusammenführung der Daten, die für die Zustandsbewertung zugrunde gelegt werden. Im zweiten Schritt prüft das technische System (RASMUS), ob ein Leistungseinbruch vorliegt und somit Adaptierungsbedarf besteht. Ist dies der Fall, untersucht es im nächsten Schritt, welche Nutzerzustände und Indikatoren kritische Ausprägungen aufweisen, die als Ursachen für den Leistungseinbruch in Betracht kommen. Im vierten Schritt bereitet RASMUS die Ergebnisse aus Schritt 2 und 3 so auf, dass sie als Entscheidungsgrundlage für die Auswahl und Anwendung von Adaptierungsstrategien (Stufe 3 und 4 der technischen Zustandsregulierung) verwendet werden können. Die folgenden Abschnitte 6.2.1 bis 6.2.4 stellen die innerhalb der einzelnen Schritte ablaufenden Prozesse näher vor.

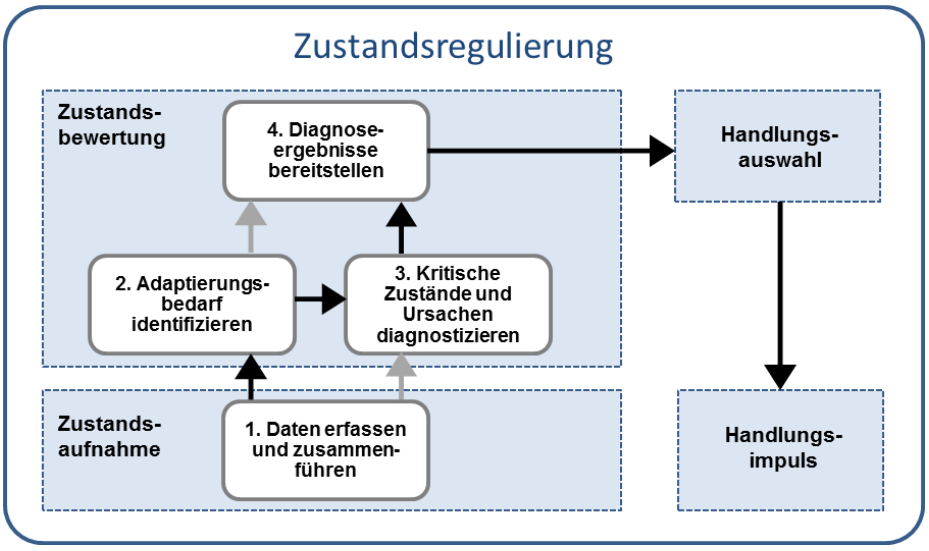


Abbildung 36. Schritte der Echtzeitdiagnose als Bestandteil der Zustandsregulierung des adaptiven Systems

6.2.1 Daten erfassen und zusammenführen (Schritt 1)

Wie in Abschnitt 6.1 beschrieben wurde, soll der Nutzerzustand sowohl multidimensional, basierend auf den sechs Zustandsdimensionen, als auch multifaktoriell bewertet werden. Dies beinhaltet die Erfassung von physiologischen und verhaltensbasierten Maßen, Merkmalen der Anforderungssituation und individuellen Faktoren. Um den Nutzerzustand zur Laufzeit bewerten zu können, ist es erforderlich, dass RASMUS diese Daten, die aus unterschiedlichen Quellen stammen (z.B. Sensordaten, Daten der Experimentalumgebung, Daten aus Fragebögen), in Echtzeit miteinander verknüpft und in Relation zueinander auswertet und bewertet. In Abbildung 37 ist der Ablauf zur Erfassung und Zusammenführung der für die Diagnose erforderlichen Daten in einem Aktivitätsdiagramm veranschaulicht. Die Aktivitäten sind unterteilt in Handlungen, die ein Mensch mit entsprechenden Fachkenntnissen – im Folgenden Experte genannt – vorab ausführt, und den im Diagnoseprozess kontinuierlich ablaufenden Aktivitäten, die von RASMUS durchgeführt werden.

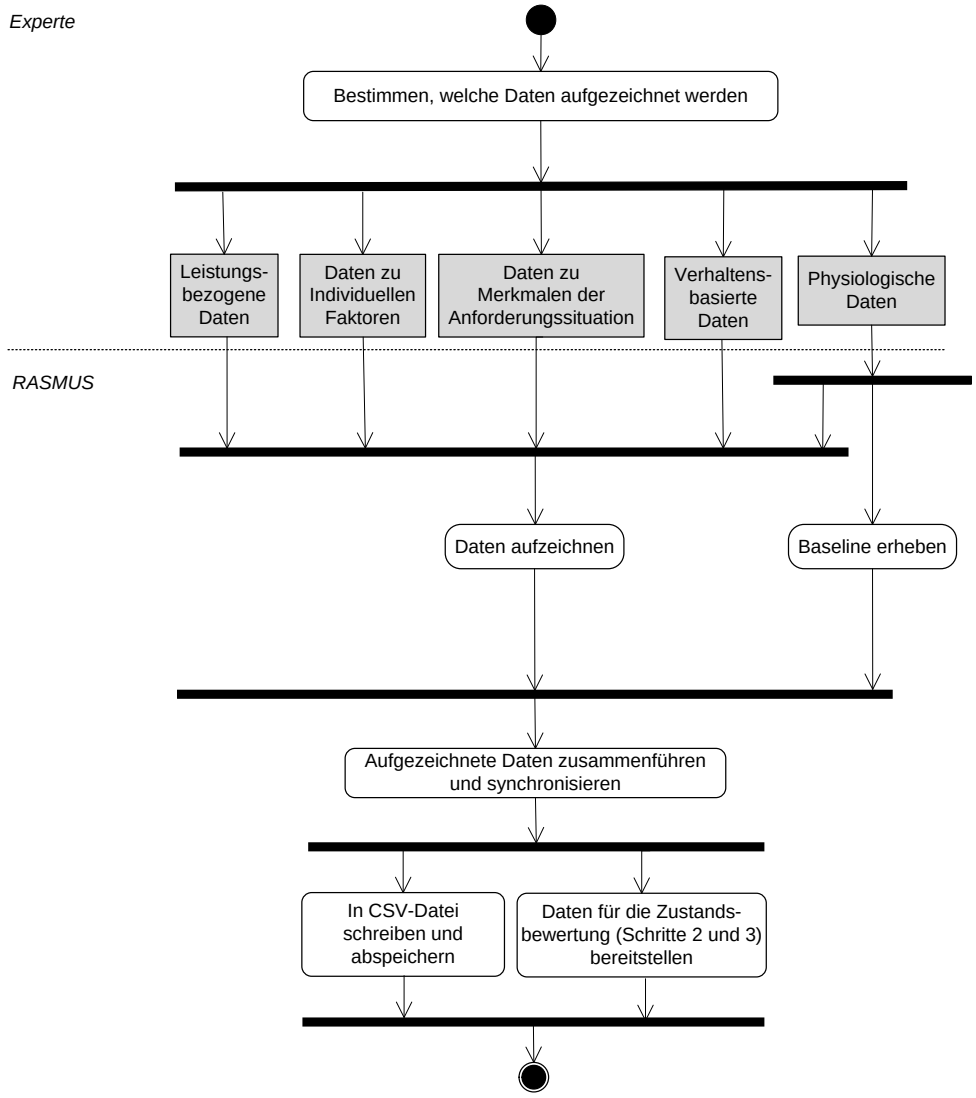


Abbildung 37. Aktivitätsdiagramm zur Erfassung und Zusammenführung der Daten (Schritt 1)

Zunächst bestimmt der Experte, welche Daten aufgezeichnet werden. Dazu zählen leistungs-bezogene Maße zur Identifizierung von Leistungseinbrüchen, sowie die bereits erwähnten Indikatoren zur multifaktoriellen Bewertung des Nutzerzustands. Die Daten werden im

Diagnoseprozess durch Sensoren und das Experimentalsystem erfasst. Bei den physiologischen Variablen ist außerdem vorab die Erhebung einer Baseline erforderlich (vgl. Anforderung in Abschnitt 2.4.3), bei der sich der Nutzer in einem noch wenig beanspruchten, leistungsfähigen Zustand befinden sollte. Die verschiedenen Datenströme werden zusammengeführt und synchronisiert. Anschließend werden die Rohdaten in eine CSV-Datei geschrieben und zur weiteren Verarbeitung an die Zustandsbewertung (Schritte 2 und 3) weitergeleitet.

6.2.2 Adaptierungsbedarf identifizieren (Schritt 2)

Nach der Erfassung und Zusammenführung der Daten ermittelt RASMUS, ob ein Leistungseinbruch besteht, der das Vorliegen von Adaptierungsbedarf anzeigt. Das Aktivitätsdiagramm in Abbildung 38 veranschaulicht diesen Prozess.

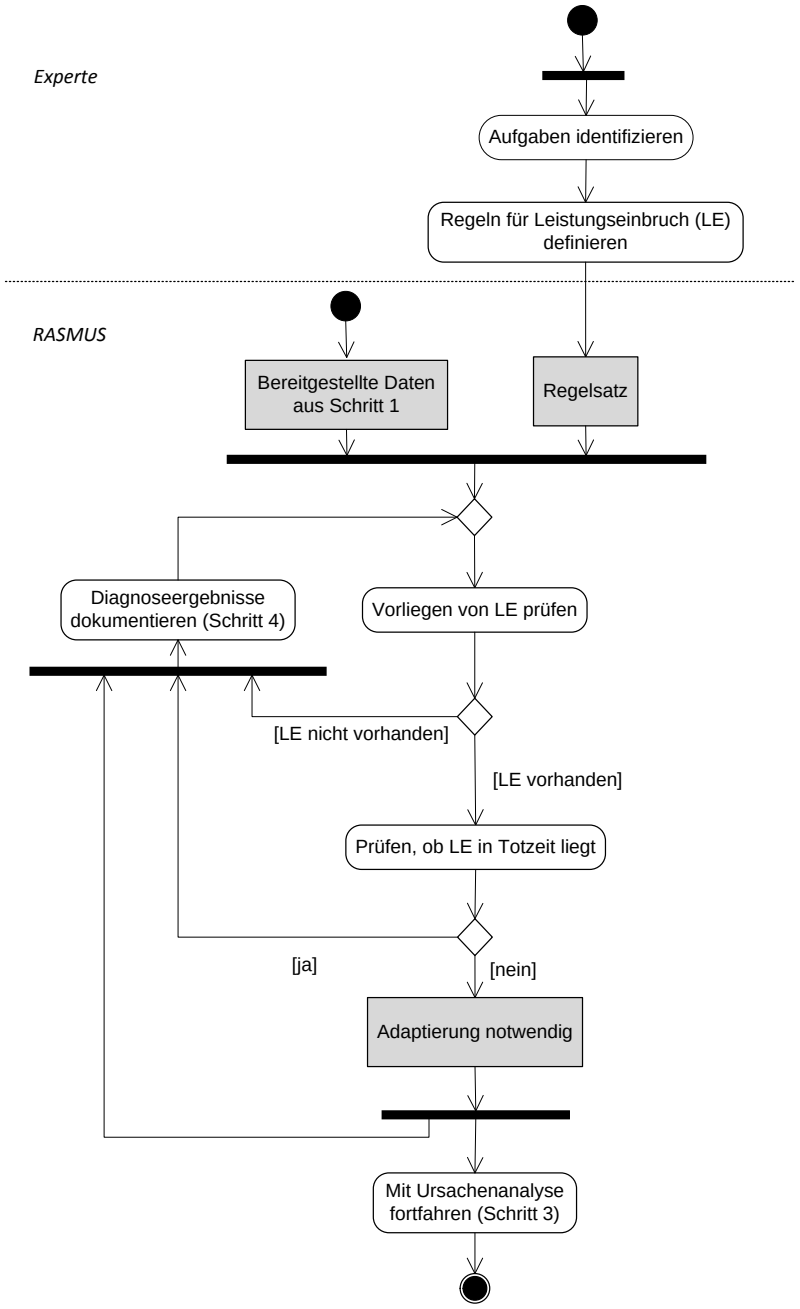


Abbildung 38. Aktivitätsdiagramm zur Bestimmung des Adaptierungsbedarfs (Schritt 2)

Der Prozess beginnt damit, dass der Experte die Aufgaben, für die Leistungseinbrüche diagnostiziert werden sollen, identifiziert und Bewertungskriterien festlegt. Mit Hilfe eines Regeleditors (siehe Abschnitt 6.7) legt der Experte die Bewertungskriterien (z.B Aufgabendauer, Fehler) und die konkreten Regeln für Leistungseinbrüche fest (z.B. Aufgabendauer > 60 Sekunden). RASMUS bewertet anhand dieser Regeln die aus Schritt 1 bereitgestellten leistungsbezogenen Daten. Wenn RASMUS für eine der Aufgaben einen Leistungseinbruch feststellt, überprüft es als nächstes, ob dieser Leistungseinbruch in der so genannten „Totzeit" liegt. Die Totzeit bezieht sich auf eine Zeitspanne nach dem letzten Leistungseinbruch, in der weitere Leistungseinbrüche keine weiteren Adaptierungen auslösen sollten. Damit soll vermieden werden, dass sich die Adaptierung mehrere Male in kurzen Zeitabständen an- und ausschaltet, da sich dies störend auf die Aufgabenbearbeitung auswirken kann (vgl. Abschnitt 2.4.7). Die Dauer der Totzeit wird von dem Experten bei der Definition von Regeln für einen Leistungseinbruch nach eigenem Ermessen festgelegt.

Sofern sich der Leistungseinbruch nicht innerhalb der Totzeit ereignet hat, diagnostiziert RASMUS eine Notwendigkeit zur Adaptierung und fährt mit der Ursachenanalyse (Schritt 3) und der Dokumentation der Diagnoseergebnisse (Schritt 4) fort. Wenn kein Leistungseinbruch oder ein Leistungseinbruch, der in der Totzeit liegt, diagnostiziert wurde, besteht kein Adaptierungsbedarf. Die Ergebnisse der Leistungsbewertung werden daraufhin lediglich dokumentiert (Schritt 4). Dieser Prozess wird sowohl bei Vorhandensein als auch bei Nicht-Vorhandensein von Adaptierungsbedarf kontinuierlich von Neuem durchlaufen.

6.2.3 Kritische Zustände und Ursachen diagnostizieren (Schritt 3)

Wenn in Schritt 2 Adaptierungsbedarf festgestellt wurde, folgt die Ursachenanalyse in Schritt 3. Dabei ist anzumerken, dass dieser Prozess in der technischen Umsetzung von RASMUS kontinuierlich durchlaufen wird, das heißt unabhängig davon, ob Adaptierungsbedarf besteht oder nicht. Die Diagnoseergebnisse der Ursachenanalyse werden allerdings erst bei Auftreten eines Leistungseinbruchs außerhalb der Totzeit dem Adaptierungsmanagement bereitgestellt und gespeichert (vgl. Schritt 4).

Der Prozess der Ursachenanalyse ist in Abbildung 39 skizziert. Zunächst legt der Experte Indikatoren und Regeln für kritische Ausprägungen der Indikatoren fest. Dazu verwendet er den in Abschnitt 6.7 beschriebenen Regeleditor. Bei den physiologischen Indikatoren dient die Abweichung von der Baseline als Bewertungskriterium. Eine kritische Ausprägung liegt dann vor, wenn die aktuellen Werte ein vorher definiertes Abweichungskriterium über- oder unterschreiten (z.B. Standardabweichung oder Prozentwert über oder unter dem Mittelwert/Median der Baseline, vgl. Abschnitt 6.6.2). Bei den übrigen Indikatoren sind absolute Grenzwerte oder Boole'sche Werte (ja/nein) als Bewertungskriterien vorgesehen. Anschließend erfolgt eine Zuordnung zu den Nutzerzustandsdimensionen, um kritische Nutzerzustände zu ermitteln. Das heißt, es wird festgelegt, bei welchen Kombinationen kritisch ausgeprägter Indikatoren ein Nutzerzustand als kritisch bewertet wird. RASMUS analysiert auf Basis dieser Regelsätze anhand der in Schritt 1 bereitgestellten Daten, welche Indikatoren des Nutzerzustands und welche Nutzerzustände kritische Ausprägungen aufweisen. Daraus generiert die Echtzeitdiagnose eine Liste kritisch ausgeprägter Indikatoren und Nutzerzustände, die als Ursache für einen Leistungseinbruch in Betracht kommen.

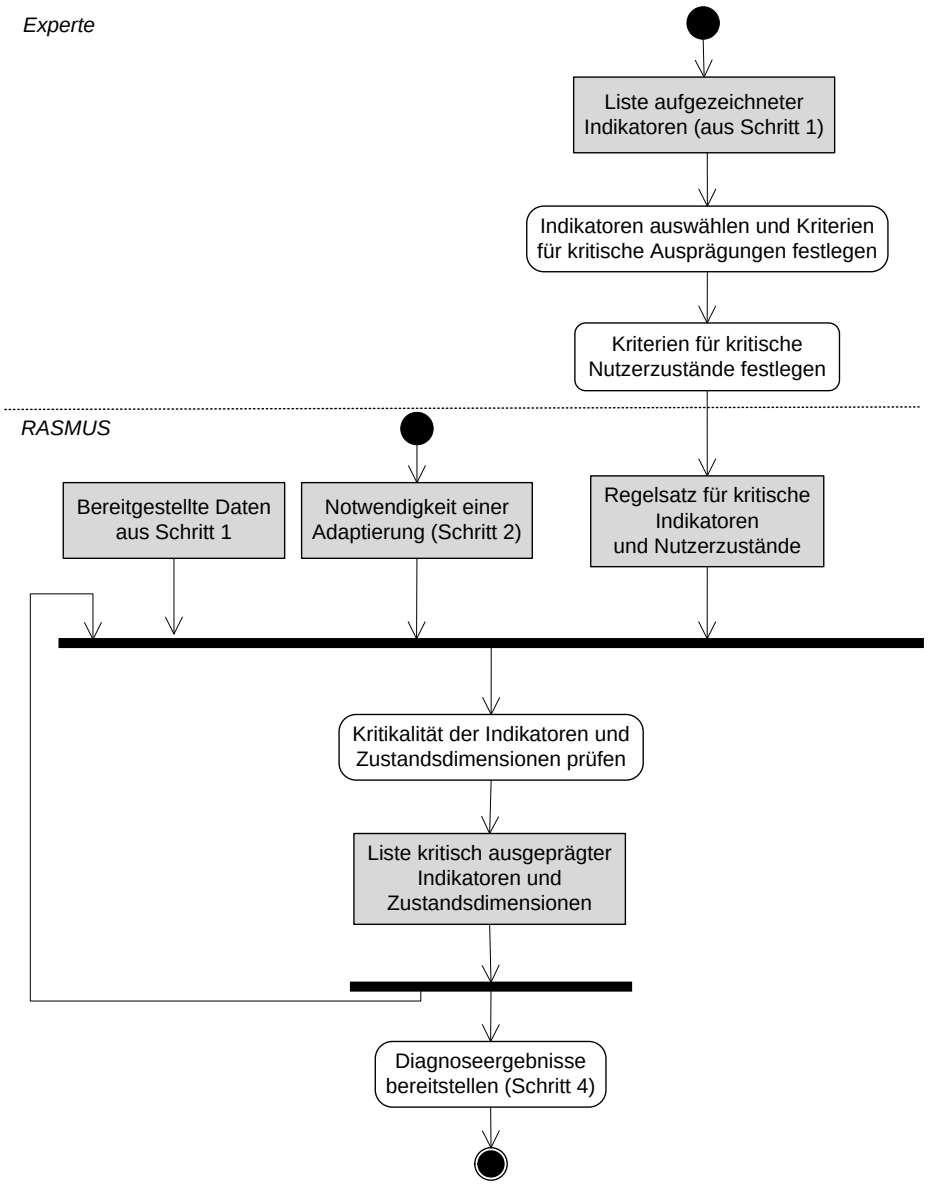


Abbildung 39. Aktivitätsdiagramm zur Ursachenanalyse (Schritt 3)

6.2.4 Diagnoseergebnisse bereitstellen (Schritt 4)

Im vierten Schritt erfolgt schließlich die Bereitstellung der Diagnoseergebnisse (vgl. Aktivitätsdiagramm in Abbildung 40). RASMUS leitet die Diagnoseergebnisse zum Adaptierungsbedarf (Schritt 2) und der Ursachenanalyse (Schritt 3) in diesem Schritt an das Adaptierungsmanagement (ADAM) weiter. Außerdem sieht das Diagnosekonzept vor, die leistungsbezogenen Daten, die Zustandsindikatoren und die daraus abgeleiteten Nutzerzustände in Echtzeit visuell darzustellen. Dies soll es Beobachtern ermöglichen, die Datenaufzeichnung zu überwachen und die Entscheidungen des adaptiven Systems nachzuvollziehen. Zur Dokumentation und für Offline-Analysen werden die Diagnoseergebnisse, die RASMUS bei Auftreten von Leistungseinbrüchen ermittelt, im CSV-Format abgespeichert.

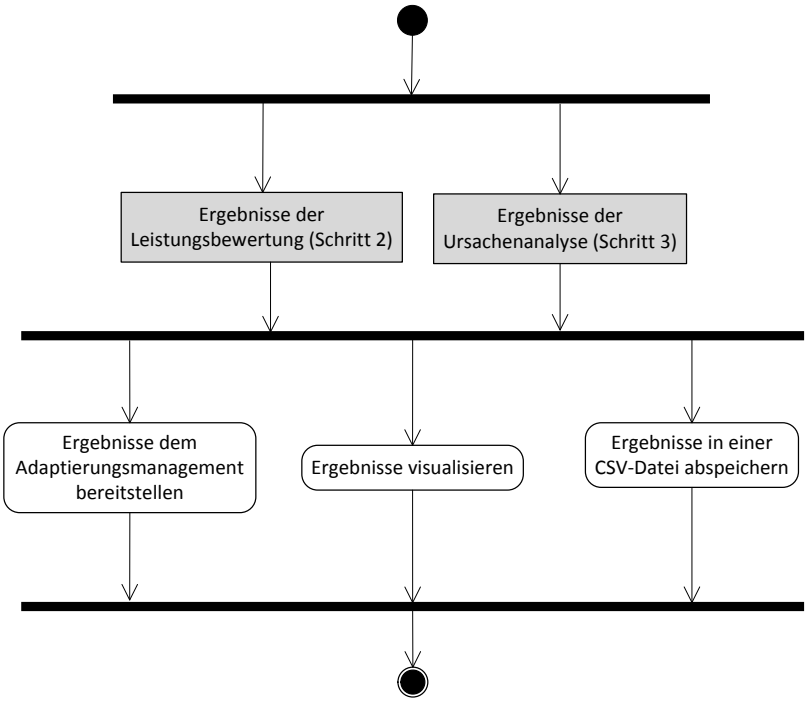


Abbildung 40. Aktivitätsdiagramm zur Bereitstellung der Diagnoseergebnisse (Schritt 4)

6.3 Experimentalparadigma für die Umsetzung

Für die Umsetzung und Validierung der Echtzeitdiagnose *RASMUS* wurde die Entscheidung getroffen, eine andere Experimentalaufgabe zu verwenden als bei den vorangegangenen experimentellen Untersuchungen, die in Kapitel 4 und 5 beschrieben sind. Wie in Abschnitt 4.5.2 bereits diskutiert wurde, kann der Szenarioverlauf bei der bisherigen Luftraumüberwachungsaufgabe in starkem Maße durch die Aktivitäten des Nutzers (Steuerung der Kontakte) beeinflusst werden. Dies kann dazu führen, dass ein Szenario trotz gleicher Ausgangsbedingungen bei verschiedenen Nutzern durch deren Interaktionen unterschiedlich verläuft und mit unterschiedlich hohen Anforderungen verbunden ist. Somit ist es schwierig, im Szenarioverlauf zielgerichtet kritische Nutzerzustände hervorzurufen, anhand derer die Validität der Diagnosefähigkeiten von RASMUS untersucht werden kann.

Bei der nun für die Umsetzung und Validierung von RASMUS verwendeten Aufgabe nimmt der Versuchsteilnehmer die Rolle eines Marineoperateurs ein, der Luftkontakte auf einem simulierten Radarbildschirm überwacht und in bestimmten Situationen Kontakte identifizieren, warnen und bekämpfen muss. Der Szenarioverlauf ist in diesem Fall genau vorgegeben und kann vom Nutzer nicht wesentlich beeinflusst werden. Für die Aufgabenbearbeitung steht ein Demonstrator mit einer grafischen Benutzungsoberfläche, kurz: GUI (Graphical User Interface), zur Verfügung, der im Auftrag der Deutschen Marine vom Fraunhofer FKIE konzipiert und prototypisch umgesetzt wurde (Kaster, Tappert, Ruckert, & Becker, 2010). Da die GUI als Gestaltungsvorlage für das Führungs- und Waffeneinsatzsystem (FüWes) der neuen Fregattenklasse F125 diente, ist hiermit, wie in Abschnitt 2.4.8 gefordert, eine realitätsnahe Aufgabenbearbeitung möglich. Die GUI ist in Abbildung 41 dargestellt (siehe auch Anhang C.2 für eine größere Ansicht).

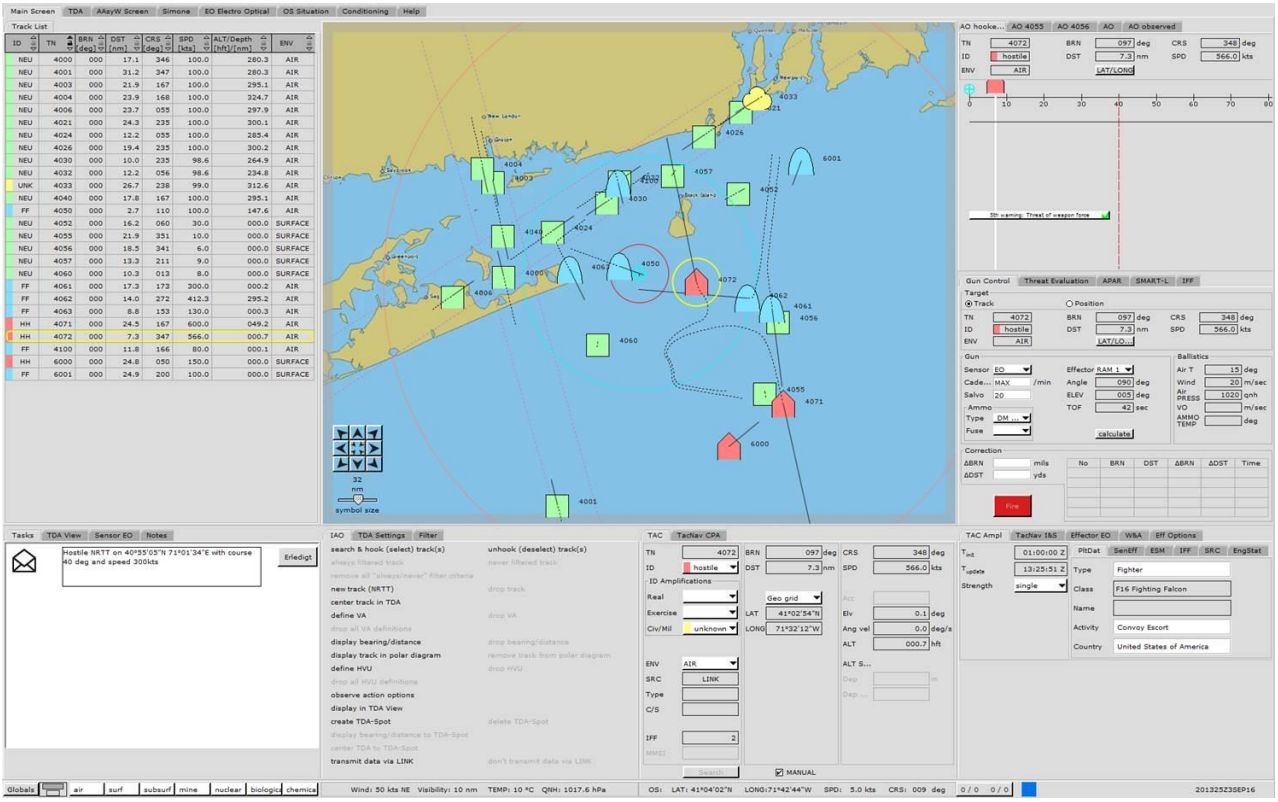


Abbildung 41. GUI zur Bearbeitung der Experimentalaufgabe

In der Mitte der GUI befindet sich die *Tactical Display Area (TDA)* – das Lagebild. Die vom Radar des Eigenschiffs erfassten Kontakte (z.B. Schiffe und Flugzeuge) sind auf der TDA als farbige Symbole dargestellt. Die Farbe zeigt dabei die Identität eines Kontakts an: gelb steht für *noch nicht identifiziert*, grün für *neutral*, blau für *freundlich*, rot für *feindlich*. Luftstraßen, auf denen sich üblicherweise zivile (neutrale) Flugzeuge befinden, sind durch gestrichelte violette Linien markiert. Das blaue Fadenkreuz im Zentrum der TDA markiert das Eigenschiff. Der rote Kreis kennzeichnet die *Weapon Range* (Waffenreichweite, WR) des Schiffs und der etwas größere blaue Kreis stellt die *Identification Safety Range* (ISR) dar. Dabei handelt es sich um eine Sicherheitszone, die in Hinblick auf unbekannte oder feindliche Kontakte zu kontrollieren ist.

Die Experimentalaufgabe beinhaltet insgesamt vier verschiedene Teilaufgaben aus dem Bereich „Anti-Air-Warfare“ (AAW), die bereits für ein früheres Experiment definiert und mit Fachpersonal der Marine abgestimmt wurden (vgl. Beschreibung des Experiments in Abschnitt 6.5). Dabei handelt es sich um das *Identifizieren, Warnen* und *Bekämpfen* von Luftkontakten sowie das *Anlegen von Non Real-Time Tracks (NRTT)* auf der TDA. NRTT sind Kontakte, die manuell auf dem Lagebild angelegt werden, da sie vom Radar nicht erfasst wurden. Eine kurze Beschreibung der Aufgaben ist Tabelle 39 zu entnehmen.

Die für die Bearbeitung dieser Aufgaben notwendigen Eingaben können in den um das Lagebild herum befindlichen Bedien- und Anzeigeelementen vorgenommen werden (siehe genauere Beschreibung in Anhang C.2). Des Weiteren befindet sich im Segment links unten (vgl. Abbildung 41) ein Nachrichtenfenster, in dem Meldungen für das Anlegen von NRTT in Form eines verschlossenen Briefumschlags angezeigt werden. Nach einem Mausklick auf den Briefumschlag

öffnet sich dieser, woraufhin die für das Anlegen des NRTT benötigten Informationen (Identität, Koordinaten, Geschwindigkeit, Höhe, Flugrichtung) angezeigt werden.

Die Aufgaben treten während des Szenarios zu unterschiedlichen Zeitpunkten und unterschiedlich häufig auf. Für den Fall, dass mehrere Aufgaben zur gleichen Zeit bearbeitet werden müssen, ist es erforderlich, die Aufgabe mit der höchsten Priorität (s. Tabelle 39 rechte Spalte) zuerst zu bearbeiten. Je höher der Zahlenwert, desto höher ist die Priorität. Das Bekämpfen ist somit die wichtigste Aufgabe mit der höchsten Priorität.

Tabelle 39. Beschreibung der in der Experimentalaufgabe verwendeten Teilaufgaben und Angabe ihrer Prioritäten

Aufgabe	Beschreibung	Priorität
Identifizieren	Noch nicht identifizierte Kontakte müssen anhand von ID-Kriterien als freundlich, feindlich oder neutral identifiziert werden. Ändern bereits identifizierte Kontakte ihr Verhalten, so dass dies eine Änderung ihrer Identität notwendig macht, müssen diese umidentifiziert werden.	100 (außerhalb der ISR) 300 (innerhalb der ISR)
Anlegen von NRTT	Zu bestimmten Zeitpunkten innerhalb des Szenarios müssen NRTT manuell auf der TDA angelegt werden. Dies wird dem Nutzer visuell durch eine Meldung angezeigt.	200
Warnen	Kontakte, die als feindlich identifiziert wurden, müssen gewarnt werden, sobald sie in die ISR eindringen.	400
Bekämpfen	Feindliche Kontakte, die trotz Warnung auf das Eigenschiff zufliegen, müssen bekämpft werden, sobald sie in die WR eindringen.	500

6.4 Betrachtete Nutzerzustände

Für die Umsetzung der Echtzeitdiagnose RASMUS wurden drei kritische Nutzerzustände ausgewählt, anhand derer die Funktionsweise des Diagnosekonzepts demonstriert und die Validität der Diagnoseergebnisse untersucht werden soll. Dabei handelt es sich um:

- *hohe mentale Beanspruchung*
- *passive aufgabenbezogene Müdigkeit*
- *falsche Aufmerksamkeitsverteilung*

Diese Nutzerzustände wurden ausgewählt, da sie in Gesprächen mit Anwendern der Bundeswehr als besonders praxisrelevant für Aufgaben im Bereich der Luftraumüberwachung eingestuft wurden. Wichtig bei der Auswahl war außerdem, dass die Nutzerzustände experimentell modulierbar sind. So sollten sich die betrachteten kritischen Zustände im Validierungsexperiment durch Veränderungen im Szenario möglichst zuverlässig hervorrufen lassen, um überprüfen zu können, ob RASMUS in der Lage ist, diese Zustände zu erkennen.

Für *hohe mentale Beanspruchung* wird auf Basis der Erkenntnisse aus Abschnitt 3.1.1 angenommen, dass dieser Zustand durch hohe mentale Belastung ausgelöst wird und mit hohem Arousal assoziiert ist. *Passive aufgabenbezogene Müdigkeit* wird demgegenüber durch länger andauernde Phasen von Monotonie und geringer Belastung ausgelöst und ist durch niedriges Arousal charakterisiert (vgl. Abschnitt 3.1.4). Da beide Zustände gegensätzliche Eigenschaften

aufweisen, kann davon ausgegangen werden, dass sie nicht gemeinsam auftreten können. Bei der *falschen Aufmerksamkeitsverteilung* handelt es sich um einen Zustand, bei dem der Nutzer, z.B. aufgrund von stressbedingtem „attentional tunneling“ (Wickens, 2005) oder müdigkeitsbedingtem Vigilanzabfall, wichtige Informationen oder Aufgaben mit hoher Priorität nicht oder ungenügend wahrnimmt (vgl. Abschnitt 3.1.5). Dieser Zustand kann somit durch hohe Beanspruchung und passive aufgabenbezogene Müdigkeit gleichermaßen ausgelöst werden.

6.5 Experiment 3 – Post hoc-Analyse zur Umsetzung einer Echtzeitdiagnose

In einem vorangegangenen Experiment (vgl. Schwarz et al., 2012; Schwarz, 2013) wurden die Zustände *Overload* und *Underload*, die den Beschreibungen von Gimeno et al. (2006) zufolge mit hoher mentaler Beanspruchung und passiver aufgabenbezogener Müdigkeit vergleichbar sind, bereits anhand der in Abschnitt 6.3 beschriebenen Aufgabe experimentell moduliert und erfasst. Das Experiment war ursprünglich mit dem Ziel durchgeführt worden, die Eignung von okulomotorischen Maßen zur Erfassung von Overload- und Underload-Zuständen zu evaluieren. Im Rahmen der vorliegenden Promotionsarbeit wurden die Experimentaldaten post hoc noch einmal analysiert, um Erkenntnisse zu geeigneten Indikatoren und Regeln der multifaktoriellen Echtzeitdiagnose für die zugrunde liegende Experimentalaufgabe abzuleiten. Das methodische Vorgehen im Experiment sowie die für die Regelerstellung relevanten Ergebnisse werden nachfolgend näher erläutert.

6.5.1 Versuchsaufbau

Das Experiment wurde in Deutschland und in Italien mit jeweils 10 Marineoperateuren durchgeführt. Aufgrund von technischen Problemen bei der Durchführung der Untersuchung in Deutschland wird nachfolgend jedoch nur auf die in Italien durchgeführte Untersuchung Bezug genommen.

Die Untersuchung wurde mit dem gleichen Experimentalsystem und den gleichen Aufgaben durchgeführt, die auch für die Umsetzung und Validierung der Echtzeitdiagnose verwendet wurden (vgl. Beschreibung in Abschnitt 6.3). Geringfügige Unterschiede ergeben sich daraus, dass es sich bei den Teilnehmern um Experten in diesem Aufgabenbereich handelte, für die eine möglichst realitätsnahe Aufgabenbearbeitung gewährleistet werden sollte (siehe Instruktionen zu den Aufgaben und der Bedienung des Experimentalsystems in Anhang C.1 und C.2).

Mit Hilfe der Simulationssoftware STAGE von Presagis wurden drei Szenarien von jeweils 15 Minuten Dauer mit unterschiedlichem Belastungsgrad entwickelt. Das sogenannte *Underload-Szenario* sollte eine geringe mentale Belastung, das *Normal Load-Szenario* eine mittlere und das *Overload-Szenario* eine hohe Belastung erzeugen. Für die Entwicklung der Szenarien wurde das Cognitive Task Load (CTL)-Modell von Neerincx (2003, vgl. Abschnitt 3.1.1) zugrundegelegt. Die drei Dimensionen des CTL-Modells (beschäftigte Zeit, Niveau der Informationsverarbeitung und Aufgabenwechsel) wurden in ähnlicher Weise wie bei DeGreef & Arciszewski (2009) über das *Aufgabenvolumen* (Anzahl zu bearbeitender Aufgaben), die *Komplexität der Situation* (Anzahl an Kontakten mit widersprüchlichen Informationen) und die *Variation der Aufgabenarten* moduliert.

Wie in Tabelle 40 dargestellt ist, nimmt die Anzahl der während eines Szenarios zu bearbeitenden Aufgaben auf höheren Belastungsstufen zu. Außerdem vermindert sich der Zeitabstand zwischen den Aufgaben: Im Underload-Szenario liegt zwischen den Aufgaben ein Abstand von mindestens 60 Sekunden. Im Normal Load-Szenario folgen die Aufgaben in einem Abstand von 20-60 Sekunden aufeinander. Und in der Overload-Bedingung müssen zumeist zwei bis drei Aufgaben zur gleichen Zeit bearbeitet werden.

Während im Underload-Szenario auf Kontakte mit widersprüchlichen Informationen – z.B. Kontakt folgt der Luftstraße (Kriterium für neutrale Identität) aber hält die Vorgaben zu Geschwindigkeit oder Höhe nicht ein (widerspricht dem Kriterium für neutrale Identität, vgl. ID-Kriterien in Anhang C.1) – verzichtet wurde, treten im Normal Load-Szenario zwei Kontakte und im Overload-Szenario fünf Kontakte mit widersprüchlichen Informationen auf. Hinsichtlich der Aufgabenarten beschränkt sich das Underload-Szenario auf das Identifizieren von Kontakten und das Anlegen eines NRTT. Im Normal Load-Szenario treten, wie auch im Overload-Szenario, alle vier Aufgabenarten auf, wobei Warnungen und Bekämpfungen im Vergleich zum Overload-Szenario deutlich seltener erforderlich sind (vgl. Tabelle 40).

Tabelle 40. Anzahl der Teilaufgaben pro Versuchsszenario (Belastungsstufe) – Experiment 3

	Identifizieren	NRTT	Warnen	Bekämpfen	Gesamt
Underload	5	1	0	0	6
Normal Load	13	3	3	3	22
Overload	23	6	12	11	52

Jeder Teilnehmer bearbeitete alle drei Versuchsszenarien (einfaktorielles Messwiederholungs-design). Um Reihenfolge- und Übungseffekte zu kontrollieren, wurde die Reihenfolge der Bearbeitung zwischen den Teilnehmern, wie in Tabelle 41 dargestellt, systematisch variiert.

Tabelle 41. Versuchsdesign (Experiment 3)

Versuchs-personen (VP)	Versuchsbedingung		
	Underload	Normal Load	Overload
VP 1	1	2	3
VP 2	2	3	1
VP 3	3	2	1
VP 4	2	1	3
VP 5	3	1	2
VP 6	1	2	3
VP 7	2	3	1
VP 8	3	2	1
VP 9	2	1	3
VP 10	3	1	2

Anmerkung: Nummerierung entspricht der Reihenfolge, in der die Szenarien von den Versuchspersonen bearbeitet wurden.

6.5.2 Abhängige Variablen

Über den Eyetracker Tobii X120 wurden während der Szenariobearbeitung okulomotorische Maße aufgezeichnet, die im Rahmen der damaligen Zielsetzung des Experiments als Indikatoren der mentalen Beanspruchung untersucht wurden (vgl. Schwarz, 2013). Im Folgenden werden die *Pupillenweite* und die *Fixationsdauer* näher betrachtet, die bereits in den Experimenten 1 und 2

(vgl. Kapitel 4 und 5) untersucht wurden und sich dort als vielversprechende Indikatoren der mentalen Beanspruchung herausgestellt hatten. Des Weiteren soll mit der *Anzahl Mausklicks* eine verhaltensbasierte Variable untersucht werden. Sie wurde bislang kaum in Studien zur mentalen Beanspruchung berücksichtigt. Aufgrund von Beobachtungen während des Experiments wird jedoch vermutet, dass die Anzahl Mausklicks ein Indikator für Underload- und Overload-Zustände in dem betrachteten Aufgabenbereich sein könnte. Dies soll daher anhand der vom Experimentalsystem aufgezeichneten Daten zum Auftreten von Mausklicks geprüft werden.

Um zu verifizieren, dass die Belastungsstufen mit unterschiedlichen Beanspruchungszuständen korrespondieren, wurde die mentale Beanspruchung außerdem subjektiv nach jedem Szenario über den Fragebogen *NASA-TLX* bewertet (vgl. Abschnitt 4.2.6). Da an der Studie sowohl deutsche als auch italienische Operateure teilnahmen, wurde aus Vergleichbarkeitsgründen die englischsprachige Originalversion verwendet.

6.5.3 Hypothesen und Forschungsfragen

Das Ziel der Analysen bestand darin, geeignete Indikatoren zur Erfassung von Underload-Zuständen bzw. passiver aufgabenbezogener Müdigkeit und Overload-Zuständen bzw. hoher mentaler Beanspruchung bei der neu gewählten Experimentalaufgabe zu identifizieren. Es wurde daher zunächst geprüft, ob sich die betrachteten Maße hinsichtlich der verschiedenen Belastungsstufen *Underload*, *Normal Load* und *Overload* signifikant unterscheiden.

Analog zu den Hypothesen in Experiment 1 wird angenommen, dass die *subjektive Bewertung* (NASA-TLX) und die *Pupillenweite* auf einer höheren Belastungsstufe signifikant höher ausfallen als auf einer niedrigeren. Bei der *Fixationsdauer* wird erwartet, dass diese mit zunehmender Belastung abnimmt. Für die *Anzahl Mausklicks* wird aufgrund von Beobachtungen angenommen, dass eine höhere Aufgabenbelastung mit einer höheren Anzahl an Mausklicks einhergeht. Die statistischen Hypothesen zu den angenommenen Unterschieden zwischen den Versuchsbedingungen (Faktorstufen) sind für jeden Indikator in Tabelle 42 dargestellt.

Tabelle 42. Angenommene Mittelwertunterschiede zwischen den Faktorstufen Underload (UL), Normal Load (NL) und Overload (OL) für die betrachteten Indikatoren in Experiment 3

Indikator	Zusammenhang mit Beanspruchung	Statistische Hypothese
NASA-TLX	positiv	$\mu_{UL} < \mu_{NL} < \mu_{OL}$
Anzahl Mausklicks	positiv	$\mu_{UL} < \mu_{NL} < \mu_{OL}$
Pupillenweite	positiv	$\mu_{UL} < \mu_{NL} < \mu_{OL}$
Fixationsdauer	negativ	$\mu_{UL} > \mu_{NL} > \mu_{OL}$

Neben der hypothesenprüfenden Untersuchung wurden die physiologischen und verhaltensbasierten Indikatoren, die sich in Hinblick auf die verwendete Experimentalaufgabe als sensitiv für Veränderungen der Belastungsstufe erweisen, im zeitlichen Verlauf näher betrachtet. Ziel war es, hierdurch Erkenntnisse zu gewinnen, welche Indikatorausprägungen auf Underload-Zustände bzw. passive aufgabenbezogene Müdigkeit und Overload-Zustände bzw. hohe mentale Beanspruchung

hinweisen. Diese Befunde dienten als Grundlage, um in den Diagnoseregeln geeignete Grenzwerte für kritische Indikatorausprägungen festzulegen (vgl. Abschnitt 6.6.2).

Das Diagnosekonzept sieht außerdem vor, Einflussfaktoren auf den Nutzerzustand in die Diagnose einzubeziehen (vgl. Abschnitt 6.1). Ein Einflussfaktor, der im Experiment systematisch zwischen den Versuchsbedingungen verändert wurde, um die Belastung zu variieren und bei den Operateuren Zustände unterschiedlich hoher Beanspruchung hervorzurufen, ist die Anzahl der zu bearbeitenden Aufgaben. Diese ist pro Szenario für jeden Teilnehmer gleich hoch (vgl. Tabelle 38). Allerdings kann die Anzahl der gleichzeitig zu bearbeitenden Aufgaben pro Zeiteinheit je nach Bearbeitungsgeschwindigkeit schwanken: Je langsamer die Aufgaben bearbeitet werden, desto höher ist die Zahl der Aufgaben, die gleichzeitig bearbeitet werden müssen und desto höher ist die Belastung, die auf den Teilnehmer zu dem jeweiligen Zeitpunkt einwirkt. Sofern keine Motivationsprobleme vorliegen (vgl. Abschnitt 3.1.3), sollte dies mit einer höheren mentalen Beanspruchung einhergehen. Die Variable *Anzahl Aufgaben* stellt damit ebenfalls einen potenziellen Indikator für Underload- und Overload-Zustände dar, der im zeitlichen Verlauf näher analysiert wird (vgl. Abschnitt 6.5.8).

6.5.4 Stichprobe und Durchführung

Die Untersuchung in Italien wurde am Marinestützpunkt in Taranto mit 10 Operateuren der Marine (alle männlich, 36-50 Jahre, *M*=39) durchgeführt. Die Teilnehmer waren im Durchschnitt seit 20 Jahren (Minimum 15 Jahre) bei der Marine und gaben an, in dem Aufgabenbereich der Luftraumüberwachung sehr gute (3 Teilnehmer) oder gute Kenntnisse (7 Teilnehmer) zu besitzen.

Der für die Untersuchung vorgesehene Raum wurde durch lichtundurchlässige Gardinen abgedunkelt und künstlich beleuchtet, um konstante Lichtbedingungen zu gewährleisten. Jede Person nahm einzeln an dem Versuch teil. Da den teilnehmenden Operateuren die Benutzungsschnittstelle unbekannt war, erhielten sie einige Tage vor ihrer Teilnahme eine schriftliche Beschreibung des Experimentalsystems mit Instruktionen zur Bedienung (siehe Anhang C.2). Der Experimentalleiter erläuterte die Art und Weise der Aufgabenbearbeitung außerdem noch einmal zu Beginn der Untersuchung am Experimentalsystem. Die Aufgaben sowie die Kriterien zur Identifizierung („IDCrits“) von Kontakten waren auf einem Instruktionsblatt (siehe Anhang C.1) beschrieben, das die Teilnehmer auch während der Aufgabenbearbeitung zu Hilfe nehmen konnten.

In einem Übungsszenario von ca. 10 Minuten Dauer konnten sich die Teilnehmer mit der Experimentalumgebung und den Aufgaben vertraut machen. Anschließend erfolgte nach einer Kalibrierung des Eyetrackers die Aufgabenbearbeitung in den drei Versuchsszenarien *Underload*, *Normal Load* und *Overload* entsprechend der im Versuchsplan vorgegebenen Reihenfolge (vgl. Tabelle 41). Nach jedem Szenario wurden die Teilnehmer gebeten, ihre Beanspruchung mit dem Fragebogen NASA-TLX zu bewerten. Dieser wurde in elektronischer Form auf dem Computer dargeboten. In der letzten Befragungsphase wurden außerdem Informationen zu Alter, Geschlecht, Erfahrung und Rang erfragt. Pro Teilnehmer betrug die Dauer der Untersuchung ca. 1 - 1,5 Stunden.

6.5.5 Datenaufbereitung und -auswertung

Die Datenaufbereitung erfolgte in ähnlicher Weise wie in Experiment 1 (vgl. Abschnitt 4.2.9). Sie beinhaltete die Bereinigung der Daten von ungültigen Werten, die Synchronisierung und z-Standardisierung. Da mit dem Eyetracker nur ein physiologischer Sensor verwendet wurde und das Auftreten von Aufgaben und Mausklicks eventbasiert geloggt wurde, konnte auf eine Aggregierung der Daten auf Sekundenbasis verzichtet werden.

Die statistische Auswertung erfolgte mit dem Statistikprogramm SPSS 20.0 (IBM Corp., 2011) und die grafische Visualisierung der Ergebnisse mit Microsoft Excel 2010. Die in Abschnitt 6.5.3 beschriebenen Hypothesen wurden nach Prüfung der Voraussetzungen für jeden Indikator über Varianzanalysen mit Messwiederholung untersucht. Die Prüfung der Varianzhomogenität bzw. Sphärizität erfolgte über den von SPSS ausgegebenen Mauchly-Test. Bei signifikantem Ergebnis wurde eine Korrektur der Freiheitsgrade nach dem Greenhouse-Geisser-Verfahren vorgenommen (Greenhouse & Geisser, 1959). Für die weitere Untersuchung der Unterschiede zwischen den einzelnen Belastungsstufen wurden Paarvergleiche durchgeführt, bei denen zur Vermeidung einer Alpha-Fehler-Kumulierung die Bonferroni-Holm-Korrektur angewandt wurde.

Für die Untersuchung der Indikatoren im zeitlichen Verlauf ist bei den physiologischen Maßen eine Glättung durch Aggregierung oder Mittelwertbildung über definierte Zeitintervalle erforderlich, um starke Oszillationen abzuschwächen (vgl. Abschnitt 2.4.7). Eine Intervallbildung ist außerdem bei diskreten Variablen notwendig, bei denen Häufigkeiten pro Zeiteinheit berechnet werden, wie zum Beispiel bei der Variable Anzahl Mausklicks. Als Intervall-Länge wurden 30 Sekunden als geeignet erachtet, welche auch in anderen Studien angewendet wurde (z.B. Roscoe, 1993; Benson, Huddleston, & Rolfe, 1965).

6.5.6 Ergebnisse

Die varianzanalytische Untersuchung ergibt für alle Indikatoren signifikante Haupteffekte hinsichtlich der pro Versuchsbedingung variierten Belastungsstufen (vgl. Tabelle 43). Für den *NASA-TLX*, die *Pupillenweite* und die *Anzahl Mausklicks* fällt der Haupteffekt mit $p<.01$ signifikant aus, entsprechend weist auch das Effektstärkemaß η_p^2 auf einen großen Effekt hin. Für die *Fixationsdauer* liegt dagegen ein deutlich schwächerer Haupteffekt vor.

Tabelle 43. Ergebnisse zu Haupteffekten und Paarvergleichen zwischen den Belastungsstufen (Experiment 3)

	Haupteffekt Bedingung			**UL vs. NL**		**NL vs. OL**		**UL vs. OL**	
	F(2,18)	*p*	η_p^2	F(1,9)	*p*	F(1,9)	*p*	F(1,9)	*p*
NASA-TLX	38.10[#]	<.001	.81	13.58	<.01	15.86	<.01	193.87	<.001
Fixationsdauer	4.89	<.05	.35	0.2	.67	7.58	<.05[a]	4.82	.056
Pupillenweite	38.89[#]	<.001	.81	11.27	.01	15.18	<.01	2038.13	<.001
Mausklicks	270.36[#]	<.001	.97	27.93	.001	217.66	<.001	5977.62	<.001

UL = Underload, NL = Normal load, OL = Overload; [#] Sphärizität nicht gegeben, Angabe der Signifikanz nach Greenhouse-Geisser-Korrektur, [a] bei Anwendung der Bonferroni-Holm-Korrektur nicht signifikant.

Aus dem in Abbildung 42 dargestellten Balkendiagramm geht hervor, inwiefern sich die Indikatoren zwischen den Versuchsbedingungen unterscheiden. Bei dem *NASA-TLX*, der *Pupillen-*

weite und der *Anzahl Mausklicks* zeigen sich große hypothesenkonforme Unterschiede zwischen den Versuchsbedingungen, die – wie aus den Paarvergleichen in Tabelle 43 hervorgeht –signifkant ausfallen. Die subjektive Beanspruchung, die Pupillenweite und die Anzahl Mausklicks sind somit in der Underload-Bedingung am geringsten und in der Overload-Bedingung am höchsten ausgeprägt. Anders verhält es sich bei der *Fixationsdauer*. Es wurde erwartet, dass sich diese mit zunehmender Belastung verringert. Es zeigt sich jedoch, dass die Fixationsdauer im Overload-Szenario höher ausfällt als im Underload und im Normal Load-Szenario. Zwischen dem Underload- und dem Normal Load-Szenario liegt kein signifikanter Unterschied vor. Die Ergebnisse fallen für die Fixationsdauer somit nicht hypothesenkonform aus.

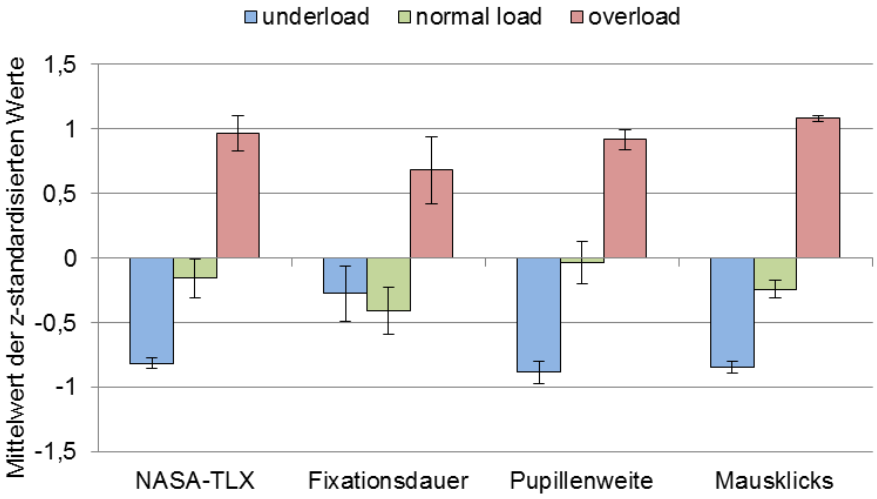


Abbildung 42. Mittelwerte und Standardfehler des NASA-TLX-Gesamtscore, der Fixationsdauer, der Pupillenweite und der Anzahl Mausklicks pro Versuchsbedingung (Experiment 3)

Da in den vorangegangenen Experimenten (vgl. Kapitel 4 und 5) deutlich wurde, dass sich die Ergebnisse interindividuell stark unterscheiden können, wurden die Ergebnisse zu den Unterschieden zwischen den Faktorstufen auch für jede Person einzeln aufbereitet. Die Diagramme sind in Anhang C.3 dargestellt. Sie zeigen, dass die *subjektive Bewertung* und die *Anzahl Mausklicks* bei nahezu allen Teilnehmern im Underload-Szenario am geringsten und im Overload-Szenario am höchsten ausfällt. Für die *Pupillenweite* gilt dies weitgehend ebenfalls. Nur bei wenigen Teilnehmern weist das Normal Load-Szenario entweder eine höhere Pupillenweite als das Overload-Szenario oder eine niedrigere Pupillenweite als das Underload-Szenario auf. Im Vergleich dazu sind die Unterschiede bei der *Fixationsdauer* zwischen den Teilnehmern deutlich stärker; teilweise treten auch gegensätzliche Ergebnisse auf. Eine klare Aussage zum Verhalten der Fixationsdauer bei unterschiedlichen Belastungsgraden kann auf Individualebene somit nicht getroffen werden.

6.5.7 Diskussion

Die Ergebnisse der subjektiven Beanspruchungsbewertung durch den NASA-TLX weisen darauf hin, dass die drei Szenarien erfolgreich unterschiedliche Beanspruchungszustände erzeugen konnten. Dies spiegelt sich auch in der *Pupillenweite* und der *Anzahl Mausklicks* wider, die somit als geeignet für eine Erfassung von Underload- und Overload-Zuständen im betrachteten Experimentalparadigma erachtet werden können. Im Unterschied zu den Befunden aus den zuvor

beschriebenen Experimenten erwies sich die *Fixationsdauer* hingegen für die neue Experimentalaufgabe als nicht geeignet, um zwischen Underload- und Overload-Zuständen zu unterscheiden. Dies könnte darauf zurückzuführen sein, dass sich auf höheren Belastungsstufen – im Unterschied zum vorhergehenden Experimentalparadigma – nicht nur die Anzahl sondern auch die Art der Aufgaben verändert hat (häufigere Warnungen und Bekämpfungen, siehe Tabelle 40). Da Literaturbefunde darauf hinweisen, dass der Zusammenhang zwischen Fixationsdauer und Aufgabenbelastung je nach Aufgabenart positiv oder negativ ausfallen kann (vgl. Abschnitt 2.3.3), könnte die unterschiedliche Häufigkeit der verschiedenen Aufgabenarten in den drei Szenarien zu den unerwarteten Effekten beigetragen haben. Darüber hinaus weist die Untersuchung auf Individualebene darauf hin, dass sich die szenariospezifischen Einflüsse auch interindividuell unterschiedlich auf die Fixationsdauer ausgewirkt haben. Es ist somit fraglich, ob die Fixationsdauer ein geeigneter Indikator des Nutzerzustands im vorliegenden Experimentalparadigma ist.

6.5.8 Untersuchung zum Verhalten der Diagnosemaße im zeitlichen Verlauf

Da die Variablen *Pupillenweite* und *Anzahl Mausklicks* erwartungskonform zwischen den verschiedenen Belastungsstufen diskriminieren konnten, werden diese, wie auch auch die Variable *Anzahl Aufgaben,* näher im zeitlichen Verlauf untersucht. In Abbildung 43 ist der Median und die Spannweite dieser drei Variablen in 30 Sekunden-Intervallen dargestellt. Das *Underload*-Szenario ist dabei blau hinterlegt, das *Normal Load*-Szenario grün und das *Overload*-Szenario rot. Da das Konzept von RASMUS vorsieht, bei den physiologischen Maßen Abweichungen von einer Baseline zu untersuchen, wurden für die Pupillenweite die z-standardisierten Werte herangezogen. Bei der Anzahl Mausklicks erscheint eine vorherige Erhebung einer Baseline nicht praktikabel, da hierfür eine längere Voraberhebung an der Experimentalaufgabe notwendig wäre. Daher wurden für diese Variable die absoluten Häufigkeitswerte im zeitlichen Verlauf untersucht. Gleiches gilt für die Anzahl Aufgaben.

Bei der *Pupillenweite* zeigt sich, dass der Median im Underload-Szenario größtenteils unterhalb und im Overload-Szenario weitgehend oberhalb des Gesamtmittelwerts von 0 verläuft. Dies bedeutet, dass auch bei Betrachtung im zeitlichen Verlauf die Mehrheit der Teilnehmer während der Aufgabenbearbeitung im Underload-Szenario eine geringere und im Overload-Szenario eine höhere Pupillenweite im Vergleich zum Gesamtmittelwert aufweisen. Im Normal Load-Szenario schwankt der Median im Bereich +/- 1 um den Nullpunkt. Dieser Bereich könnte somit als Normalbereich für unkritische Abweichungen definiert werden.

Auch bei den Variablen *Anzahl Mausklicks* weist der Medianverlauf auf höheren Belastungsstufen zumeist auch höhere Werte auf. Da bei dieser Variable keine Werte unterhalb von 0 auftreten können, die Skala nach oben jedoch offen ist, weichen die Werte der Teilnehmer allerdings in den meisten Fällen stärker nach oben vom Median ab als nach unten (vgl. Abbildung 43). Gleiches gilt für die Anzahl Aufgaben im Normal Load- und im Overload-Szenario. Im Underload-Szenario konnten die Aufgaben hingegen erwartungsgemäß von allen Teilnehmern bearbeitet werden, bevor eine neue Aufgabe aufgetreten ist, so dass maximal eine Aufgabe pro 30 Sekunden-Intervall zu bearbeiten war. Auf Basis dieser Ergebnisse wurden für diese Variablen Regeln für kritische Ausprägungen abgeleitet, die in Abschnitt 6.6.2 dargestellt werden.

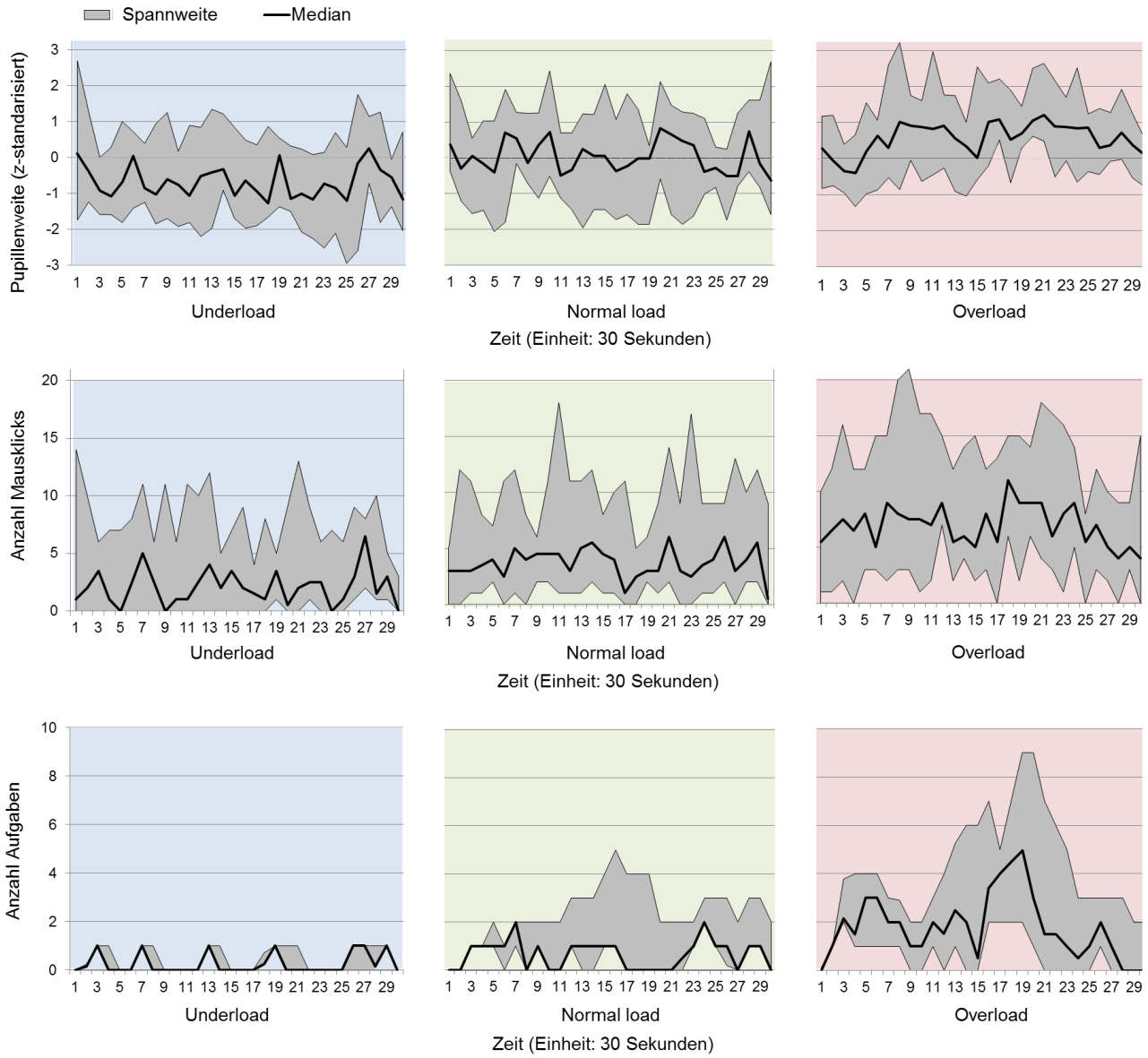


Abbildung 43. Zeitlicher Verlauf der Variablen Pupillenweite, Anzahl Mausklicks und Anzahl Aufgaben pro Szenario gemittelt über 30-Sekunden-Intervalle (Experiment 3)

6.6 Indikatoren und Regeln der Echtzeitdiagnose

In diesem Abschnitt werden die Indikatoren und Regeln zur Diagnose von Leistungseinbrüchen und kritischen Nutzerzuständen vorgestellt, die für die Echtzeitdiagnose im betrachteten Experimentalparadigma zugrunde gelegt werden sollen. Als Grundlage für die Erstellung der Regeln dienten die Ergebnisse der Post hoc-Analyse zu Experiment 3 (vgl. Abschnitt 6.5). Da es sich bei den Versuchsteilnehmern in Experiment 3 um Marineoperateure handelte, die Echtzeitdiagnose jedoch auch für Novizen ausgelegt sein sollte, wurden die Indikatoren und Regeln zusätzlich in Probedurchläufen unter Verwendung des Versuchsszenarios aus dem Validierungsexperiment (vgl. Abschnitt 7.1.3) mit Novizen getestet und weiter angepasst.

6.6.1 Indikatoren und Regeln zur Diagnose von Leistungseinbrüchen

Das Diagnosekonzept sieht vor, Adaptierungsbedarf über das Vorliegen eines Leistungseinbruchs zu bestimmen (vgl. Abschnitt 6.2.2). Als Indikatoren für einen Leistungseinbruch sollen die *Dauer* und die *Korrektheit* bei der Bearbeitung der verschiedenen Teilaufgaben herangezogen werden. Da sich die Teilaufgaben darin unterscheiden, wie zeitkritisch sie sind, und wieviel Zeit die Bearbeitung in Anspruch nimmt, erscheint es zweckmäßig, die Zeitdauer, nach deren Überschreiten RASMUS einen Leistungseinbruch diagnostiziert, für jede Aufgabe individuell festzulegen. Die pro Aufgabe festgelegten Grenzwerte sind in Tabelle 44 aufgeführt. Bei der Identifizierung wird neben der Dauer auch überprüft, ob die Identifizierung korrekt ist. Bei einer falschen ID-Vergabe hat der Nutzer die Möglichkeit, diese innerhalb von 30 Sekunden zu korrigieren, bevor ein Leistungseinbruch ausgelöst wird. Die Totzeit, also die Zeit nach einem Leistungseinbruch, in der ein weiterer Leistungseinbruch keinen Adaptierungsbedarf auslöst, wurde auf 60 Sekunden festgelegt.

Tabelle 44. Regeln zur Bestimmung von Leistungseinbrüchen (LE) pro Teilaufgabe

Aufgabe	Zeitdauer in Sek. bis zum LE
Identifizieren	60
ID-Korrektur	30
Anlegen eines NRTT	90
Warnen	20
Bekämpfen	10

6.6.2 Indikatoren und Regeln zur Diagnose der Nutzerzustände

In Tabelle 45 sind die Indikatoren und Regeln dargestellt, anhand derer RASMUS kritische Ausprägungen der betrachteten drei Nutzerzustände in Echtzeit ermittelt. Dabei ist anzumerken, dass individuelle Faktoren, wie das Alter und die Erfahrung in der praktischen Umsetzung von RASMUS nicht als Indikatoren für kritische Nutzerzustände herangezogen wurden. Dies liegt darin begründet, dass sie für die kurze Dauer der Aufgabenbearbeitung in dem geplanten Validierungsexperiment (vgl. Kapitel 7) als konstant angesehen werden können. Damit haben sie für die Diagnose intraindividueller Veränderungen im betrachteten Anwendungsfall keinen diagnostischen Wert. Dennoch ist vorgesehen, individuelle Faktoren im Rahmen einer Vorabbefragung zu erfassen, da sie als Einflussfaktoren Aufschluss über mögliche Ursachen für kritische Nutzerzustände geben können und somit für die Auswahl geeigneter Adaptierungsstrategien bedeutsam sind. Die Diagnoseregeln werden nachfolgend näher erläutert.

Tabelle 45. Indikatoren und Regeln zur Diagnose kritischer Nutzerzustände

Hohe Beanspruchung	Passive aufg. Müdigkeit	Falsche Aufmerksamkeitsverteilung
• Anzahl Aufgaben >2	• Anzahl Aufgaben < 2	• Anzahl Aufgaben > 1 und keine Bearbeitung der Aufgabe mit höchster Priorität
• Anzahl Mausklicks >10	• Anzahl Mausklicks < 3	
• HRV > 1 SD *unter* BL	• HRV > 1 SD *über* BL	• Anzahl Aufgaben = 1 und keine Bearbeitung der Aufgabe
• Pupillenweite > 1 SD *über* BL	• Pupillenweite > 1 SD *unter* BL	
• Atemfrequenz > 1 SD *über* BL	• Atemfrequenz > 1 SD *unter* BL	

Anmerkung: Bei hoher Beanspruchung und passiver aufgabenbezogene Müdigkeit müssen mindestens 3 von 5 Kriterien erfüllt sein, bei falscher Aufmerksamkeitsverteilung eine von beiden Kriterien; (SD = Standardabweichung, BL = Baseline)

Diagnoseregeln für hohe mentale Beanspruchung und passive aufgabenbezogene Müdigkeit

Den Ausführungen in Abschnitt 6.4 entsprechend werden *hohe mentale Beanspruchung* und *passive aufgabenbezogene Müdigkeit* als gegensätzliche Zustände betrachtet. Die Diagnose dieser Zustände erfolgt daher über die gleichen Indikatoren, wobei es vom Ausprägungsgrad (geringe vs. hohe Ausprägung) abhängt, ob der Indikator auf Beanspruchung oder Müdigkeit hinweist. Entsprechend den Ergebnissen aus Experiment 3 (vgl. Abschnitt 6.5) werden die *Pupillenweite* als physiologischer Indikator, die *Anzahl Mausklicks* als verhaltensbasierter Indikator und die *Anzahl Aufgaben* als Merkmal der Anforderungssituation herangezogen. Um eine höhere Robustheit zu erzielen, werden außerdem die *Atemfrequenz* und die *HRV* hinzugenommen, die sich in dem in Kapitel 5 beschriebenen Retest (Experiment 2) als sensitiv gegenüber Veränderungen der mentalen Belastung erwiesen haben. Anzumerken ist, dass diese beiden Variablen in Experiment 3 nicht untersucht werden konnten, da der BioHarness, der diese Variablen aufzeichnet, zum Zeitpunkt der Datenerhebung (diese fand vor den Experimenten 1 und 2 statt) noch nicht als Sensor zur Verfügung stand. Allerdings ist aufgrund von Literaturbefunden (z.B. Prinzel et al., 2003) davon auszugehen, dass die beiden Maße – im Unterschied zu der Fixationsdauer – aufgabenunspezifisch Veränderungen des Erregungszustands widerspiegeln. Sie sollten daher auch für das neu ausgewählte Experimentalparadigma geeignete Indikatoren für hohe Beanspruchung und passive aufgabenbezogene Müdigkeit darstellen, wobei sicherzustellen ist, dass Störvariablen (z.B. Sprechen, körperliche Aktivität) nicht interferieren (vgl. Abschnitt 2.3.3). Auf die Verwendung des Emotiv-EEG wird aufgrund der Erkenntnisse aus dem Retest verzichtet (vgl. Abschnitt 5.6).

RASMUS klassifiziert einen Zustand dann als „kritisch“, wenn mindestens drei der fünf Indikatoren auf eine kritische Ausprägung des betreffenden Nutzerzustands hinweisen. Dies schließt aus, dass hohe Beanspruchung und passive aufgabenbezogene Müdigkeit gleichzeitig als kritisch diagnostiziert werden. Die Überprüfung der Kritikalität der physiologischen Parameter erfolgt über den Vergleich mit einer Baseline (vgl. Abschnitt 6.2). Es ist vorgesehen, dass die ersten 60 Sekunden nach Szenariostart als Baseline herangezogen werden und der Mittelwert als Referenz dient. Entsprechend der in Tabelle 45 angegebenen Grenzwerte ermittelt RASMUS dann eine kritische Abweichung von der Baseline, wenn der Mittelwert der letzten 30 Sekunden (gleitender Mittelwert) eine Standardabweichung ober- und unterhalb des Mittelwerts der Baseline liegt. Die Grenzwerte für die *Anzahl Aufgaben* und die *Anzahl Mausklicks* wurden auf Basis der Erfahrungswerte aus Experiment 3 und den Vortests für das Validierungsexperiment festgelegt.

Diagnoseregeln für eine falsche Aufmerksamkeitsverteilung

Für die Diagnose einer *falschen Aufmerksamkeitsverteilung* wird die Priorität der derzeitig bearbeiteten Aufgabe als Indikator herangezogen. Die Diagnose einer „kritischen“ Aufmerksamkeit erfolgt dann, wenn die Priorität der derzeitig bearbeiteten Aufgabe kleiner als die Priorität einer der nicht bearbeiteten Aufgaben ist. Die Prioritäten wurden gemäß Tabelle 39 in Abschnitt 6.3 den Aufgaben vorab zugewiesen. Da in der Experimentalaufgabe alle Teilaufgaben an bestimmte Objekte gebunden sind (Symbole für Luftkontakte oder Briefumschlag), bestimmt RASMUS über das Objekt, das zuletzt per Mausklick angewählt wurde, die Aufgabe, mit der sich der Nutzer gerade beschäftigt. Voraussetzung für das Auslösen dieser Regel ist, dass mindestens zwei Aufgaben vorhanden sind. Während der Pilotstudie für das Validierungsexperiment (vgl. Kapitel 6)

stellte sich jedoch heraus, dass auch das Überwachen der Kontakte im Lagebild als eine Aufgabe angesehen werden muss, da die Überwachung ebenfalls die Aufmerksamkeit bindet, und das Auftreten anderer Aufgaben (z.B. Briefumschlag mit NRTT-Aufgabe) dadurch übersehen werden kann. Daher wurde die Diagnoseregel dahingehend ergänzt, dass die Aufmerksamkeit auch dann als kritisch diagnostiziert werden soll, wenn nur eine Aufgabe zu bearbeiten ist, aber derzeit keine Aufgabe bearbeitet wird.

6.7 Technische Umsetzung

Die technische Umsetzung von RASMUS basiert auf dem in Abschnitt 6.2 vorgestellten Diagnoseprozess und bezieht sich auf die Experimentalaufgabe, Nutzerzustände, und Diagnoseregeln, die in den vorigen Abschnitten 6.3 bis 6.6 vorgestellt wurden. Da die technische Realisierung der Diagnosefunktionen nicht Bestandteil der Promotionsarbeit ist, werden die wesentlichen Komponenten der Diagnoseschnittstelle an dieser Stelle nur in kurzer Form erläutert. Für eine ausführlichere Beschreibung siehe Fuchs, Schwarz & Werger (2016) sowie Werger (2016).

Das Zusammenspiel der verschiedenen Softwarekomponenten in der Diagnose ist in Anhang D.1 veranschaulicht. Die Zusammenführung und Synchronisierung der Datenströme erfolgt über die Softwareplattform *iMotions™*. Sie bietet die Möglichkeit verschiedene physiologische Sensoren (u.a. die verwendeteten Eyetracker Tobii X120 und SMI-Redn) per „plug & play“ an iMotions™ anzubinden. Darüber hinaus besteht die Möglichkeit, Sensoren, für die kein Plug-in vorhanden ist (z.B. BioHarness), mit Hilfe eines SDK anzubinden sowie Aufzeichnungen des für die Aufgabensimulation verwendeten Demonstrators (u.a. Mausklicks sowie Anzahl und Art aktiver Aufgaben) über eine „External Event API“ zu integrieren. iMotions™ speichert diese Daten und leitet alle Datenströme synchronisiert an das Diagnosemodul von RASMUS weiter.

Das Diagnosemodul wertet die Daten anhand zuvor definierter Regelsätze in Echtzeit aus, um Leistungseinbrüche, kritische Indikatorausprägungen und kritische Nutzerzustände zu bestimmen. Für die Erstellung der Regelsätze steht ein Regeleditor zur Verfügung, der im Rahmen einer Bachelorarbeit konzipiert und implementiert wurde (vgl. Werger, 2016). Der Regeleditor bietet die Möglichkeit, Regeln zu Leistungseinbrüchen, kritischen Indikatorausprägungen sowie kritischen Nutzerzuständen über Eingabemasken zu erstellen, ohne dass Veränderungen an den Quellcodes der Experimentalumgebung oder der Diagnosekomponente notwendig sind (vgl. Abbildung in Anhang D.2). Erstellte Regelsätze können gespeichert und wieder geladen werden, um sie zu einem späteren Zeitpunkt zu erweitern oder zu ändern. Zu Beginn jeder Datenaufzeichnung fragt das Diagnose-modul ab, welcher der gespeicherten Regelsätze für die Diagnose zugrunde gelegt werden soll.

Nach dem Start der Datenaufzeichnung werden die Diagnoseergebnisse kontinuierlich und in Echtzeit in einem Leistungs- und Zustandsmonitor visualisiert (siehe Abbildung in Anhang D.3). Dieser zeigt an, welche und wie viele Aufgaben zu bearbeiten sind, und wie viel Zeit für die Bearbeitung einer Aufgabe zur Verfügung steht, bis sie einen Leistungseinbruch auslöst. Aufgaben, für die ein Leistungseinbruch vorliegt, sowie potenziell kritische Indikatorausprägungen und Nutzerzustände werden rot und unkritische Zustände und Ausprägungen grün dargestellt. Somit kann festgestellt werden, welche Nutzerzustände bei Vorliegen eines Leistungseinbruchs kritisch ausgeprägt sind, sowie welche Indikatoren auf einen kritischen Zustand hinweisen.

7 Experiment 4 – Validierung der multifaktoriellen Echtzeitdiagnose RASMUS

Das Validierungsexperiment wurde mit dem Ziel durchgeführt, das Konzept und die Umsetzung der multifaktoriellen Echtzeitdiagnose RASMUS empirisch zu testen. Insbesondere sollte überprüft werden, ob RASMUS in der Lage ist, bei Auftreten von Leistungseinbrüchen die drei Nutzerzustände *hohe Beanspruchung*, *passive aufgabenbezogene Müdigkeit* und *falsche Aufmerksamkeitsverteilung* korrekt zu diagnostizieren (sofern sie in Verbindung mit einem Leistungseinbruch auftreten, werden sie im Folgenden auch als *kritische Beanspruchung*, *kritische Müdigkeit* und *kritische Aufmerksamkeit* bezeichnet). Zusätzlich zu einer hypothesenprüfenden Untersuchung wird außerdem die Güte der diagnostischen Entscheidungen in einer Post.hoc-Analyse bewertet.

7.1 Methodische Umsetzung

7.1.1 Experimentalumgebung

Die Experimentalaufgabe umfasst die Bearbeitung der in Abschnitt 6.3 beschriebenen Aufgaben zur Luftraumüberwachung (AAW) auf Schiffen der Deutschen Marine. Die Teilaufgaben und die Priorisierung der Aufgaben wurden aus Experiment 3 übernommen (Abschnitt 6.4). Die Kriterien zur Identifizierung der Kontakte (vgl. Anhang E.1) sowie die erforderlichen Schritte zur Durchführung von Bekämpfungen und Warnungen wurden jedoch im Vergleich zum vorangegangenen Experiment vereinfacht, um es auch Teilnehmern ohne Vorkenntnissen in diesem Bereich zu ermöglichen, die Bearbeitung der Aufgaben schnell zu erlernen.

Die Aufgaben wurden an einem 24-Zoll-Bildschirm mit Maus und Tastatur bearbeitet (s. Abbildung 44a). Bei diesem Experiment wurde der Eyetracker *SMI REDn* verwendet, der sich unterhalb des Monitors befindet. Oberhalb des Monitors ist eine Webcam platziert. Der Zephyr BioHarness3 (in Abbildung 44a links) wurde während des Versuchs um die Brust getragen. Abbildung 44b zeigt die räumliche Anordnung von Versuchsleiterarbeitsplatz (links) und Probandenarbeitsplatz (rechts).

Abbildung 44. a: Versuchsaufbau mit Probandenarbeitsplatz; b: Arbeitsplätze des Versuchsleiters und des Probanden (Experiment 4)

Der Versuchsleiter beobachtet auf dem Leistungs- und Zustandsmonitor während der Aufgabenbearbeitung Veränderungen in der Leistung und im Nutzerzustand des Teilnehmers (vgl. Abschnitt 6.7 und Abbildung in Anhang D.3). Außerdem erhält der Versuchsleiter über die Benutzungsoberfläche von *iMotions*^{TM} Informationen und Visualisierungen zu den aufgezeichneten Daten (z.B. Blickbewegungen, Verlauf der Pupillenweite). Über die iMotions^{TM}-Komponente *FACET* ist außerdem eine Analyse des emotionalen Zustands unter Verwendung der Webcam möglich. Die Analyseergebnisse wurden in Hinblick auf zukünftige Erweiterungen der Echtzeitdiagnose mit aufgezeichnet (s. Abschnitt 8.2.1.).

7.1.2 Versuchsdesign

Das Versuchsszenario ist darauf ausgelegt, dass während der Aufgabenbearbeitung Leistungseinbrüche aufgrund kritischer Beanspruchung, kritischer Müdigkeit und kritischer Aufmerksamkeit auftreten sollten (siehe nähere Beschreibung in Abschnitt 7.1.3). Das Diagnoseergebnis *kritische Aufmerksamkeit* kann, wie in Abschnitt 6.6.2 beschrieben wurde, sowohl einzeln als auch in Kombination mit *kritischer Beanspruchung* oder *kritischer Müdigkeit* vorliegen. Beanspruchung und Müdigkeit können dagegen nicht gleichzeitig als kritisch diagnostiziert werden. Bei einem Leistungseinbruch sind somit sechs verschiedene Diagnoseergebnisse möglich (vgl. Tabelle 46), die im Versuchsdesign die unabhängigen Variablen darstellen. Zur Validierung der Diagnoseergebnisse wurden aus den in Abschnitt 2.3.2 beschriebenen Gründen subjektive Bewertungen des Nutzerzustands als Vergleichsmaße herangezogen. Die subjektiven Bewertungen (vgl. Abschnitt 7.1.4) wurden zum Zeitpunkt der Leistungseinbrüche erfasst. Um interindividuelle Unterschiede z.B. aufgrund von Antworttendenzen zu kontrollieren, wurden für die Analysen die Abweichungen von einer vorher aufgezeichneten „Baseline-Bewertung“ (vgl. Abschnitt 7.1.3) herangezogen. Es kann somit untersucht werden, ob sich die von RASMUS diagnostizierten kritischen Nutzerzustände erwartungskonform in einer vom Baselinezustand abweichenden subjektiven Bewertung widerspiegeln. Die konkreten Hypothesen hierzu werden in Abschnitt 7.1.5 beschrieben. Eine Besonderheit ergibt sich daraus, dass die Häufigkeit, mit der ein bestimmter Zustand als kritisch diagnostiziert wird, nicht beeinflusst werden kann. Die Fallzahlen können daher sowohl pro Person als auch pro diagnostiziertem Nutzerzustand unterschiedlich hoch ausfallen. Das Versuchsdesign ist somit als quasiexperimentell einzuordnen, was bei der Auswertung und Interpretation der Ergebnisse zu berücksichtigen ist (vgl. Abschnitte 7.1.8 und 7.3.6).

Tabelle 46. Darstellung des quasiexperimentellen Versuchsdesigns mit den möglichen Diagnoseergebnissen von RASMUS bei Leistungseinbrüchen (LE) und der subjektiven Bewertung des Nutzerzustands als Vergleichsmaß

		Beanspruchung kritisch	**Müdigkeit kritisch**	**Beanspruchung und Müdigkeit unkritisch**
Aufmerksamkeit	**kritisch**	LE_1 … LE_n	LE_1 … LE_n	LE_1 … LE_n
	unkritisch	LE_1 … LE_n	LE_1 … LE_n	LE_1 … LE_n

Vergleichsmaß:
Subjektive Bewertung des Nutzerzustands (Abweichung von der Baseline)

7.1.3 Versuchsaufbau

Zum Zweck der Validierung der Echtzeitdiagnose wurde bei Detektion eines Leistungseinbruchs, sofern dieser nicht in der „Totzeit“ liegt (vgl. Abschnitt 6.2.2), das Szenario angehalten und der Nutzer wurde mittels Fragebogen zu seinem aktuellen Nutzerzustand befragt. Für jeden Leistungseinbruch protokolliert das Experimentalsystem die von RASMUS diagnostizierten kritischen Nutzerzustände und Indikatoren. Um individuelle Unterschiede kontrollieren und aufgabenbezogene Veränderungen des Nutzerzustands identifizieren zu können, wird sowohl bei den physiologischen Indikatoren als auch bei den subjektiven Maßen eine Baseline erfasst, die den unkritischen Zustand widerspiegelt. Die Baselineerhebung für die subjektiven Maße erfolgte im Versuch vor Beginn sowie am Ende des Versuchsszenarios. Um ein robustes Ergebnis zu erhalten, wurde der Mittelwert aus beiden Befragungen gebildet.

Da RASMUS in der Lage sein sollte, Veränderungen im Nutzerzustand im zeitlichen Verlauf festzustellen, beinhaltet das Experiment keine voneinander abgegrenzten Versuchsbedingungen wie das vorherige Experiment 3 (vgl. Abschnitt 6.5). Das Experiment beginnt mit einer 10-minütigen Trainingsphase. Die Bearbeitung eines durchgehenden Versuchsszenarios von 45min Dauer besteht, wie in Abbildung 45 dargestellt ist, aus drei ineinander übergehenden Phasen, in denen das Auftreten der drei kritischen Nutzerzustände (hohe Beanspruchung, passive aufgabenbezogene Müdigkeit, falsche Aufmerksamkeitsverteilung) provoziert werden sollte.

Nach einer 5-minütigen *Baseline-Phase*, die zur Orientierung und Aufzeichnung der Baseline für die physiologischen Maße dient, folgt eine 10-minütige *Phase hoher Belastung*, die durch viele zur gleichen Zeit und in kurzen Intervallen zu bearbeitende Aufgaben charakterisiert ist. Als Grundlage für die Gestaltung diente das Overload-Szenario aus Experiment 3 (vgl. Abschnitt 6.5.1).

Diese Phase geht über in eine *Monotonie-Phase*, in der passive aufgabenbezogene Müdigkeit durch eine geringere Anzahl an Aufgaben und längere Intervalle zwischen den Aufgaben begünstigt werden sollte. Die Monotonie-Phase ähnelt dem Underload-Szenario aus Experiment 3. Allerdings wurde die Zeitdauer dieser Phase im Vergleich zur Underload-Bedingung von 15 Minuten auf 30 Minuten erhöht. Diese Entscheidung beruht auf Erkenntnissen aus Abschnitt 3.1.4, wonach eine längere Zeitdauer notwendig ist, um Leistungseinbußen aufgrund von passiver aufgabenbezogener Müdigkeit zu erzeugen.

Kritische Aufmerksamkeit kann phasenunabhängig und in Kombination mit hoher Beanspruchung oder passiver aufgabenbezogener Müdigkeit auftreten. Sie wird im Experiment durch das gleichzeitige Auftreten verschiedener Aufgaben mit unterschiedlicher Priorität hervorgerufen.

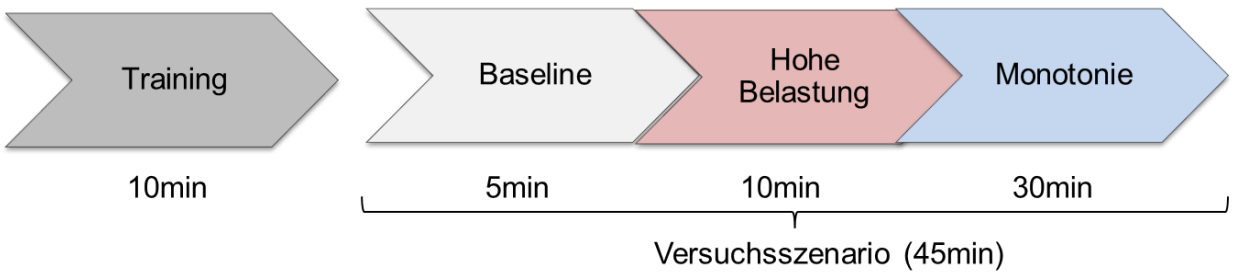


Abbildung 45. Phasen im Versuch (Experiment 4)

7.1.4 Vergleichsmaße

Zur Validierung der Diagnoseergebnisse von RASMUS wurden subjektive Bewertungen hinsichtlich der Zustandsdimensionen *Beanspruchung*, *Müdigkeit* und *Aufmerksamkeit* herangezogen. Um darüber hinaus untersuchen zu können, inwiefern andere Nutzerzustände, die (noch) nicht von RASMUS diagnostiziert werden, die Leistung beeinflusst haben, wurden diese ebenfalls subjektiv bewertet.

In der Befragung, die im Falle eines Leistungseinbruchs erfolgte, sollte der Proband seinen aktuellen Zustand somit in Hinblick auf alle sechs Dimensionen des Nutzerzustands (Beanspruchung, Aufmerksamkeit, Müdigkeit, Motivation, emotionaler Zustand, Situationsbewusstsein) bewerten. Damit die Befragung den Versuchsteilnehmer möglichst wenig bei der Aufgabenbearbeitung beeinträchtigt, wurde darauf geachtet, die Anzahl an Fragen kurz zu halten. Außerdem sollten die Fragen auf einer einheitlichen Skala beantwortet werden, um die Ergebnisse vergleichbar zu halten und für den Nutzer eine konsistente Form der Beantwortung zu gewährleisten. Ein validierter Fragebogen, der alle sechs Nutzerzustandsdimensionen abdeckt, ist derzeit nach eigenem Kenntnisstand noch nicht verfügbar. Für die Erfassung der Dimensionen des Nutzerzustands wurden daher Fragen aus bestehenden Fragebögen ausgewählt oder, sofern nicht in geeigneter Form verfügbar, selbst generiert (s. Tabelle 47). Für alle Fragen mit Ausnahme des SAM wurde die bereits in Experiment 1 (Kapitel 4) für den NASA-TLX eingesetzte 15 stufige KU-Skala von Heller (1982) als Antwortskala verwendet (vgl. Abbildung 46).

Tabelle 47. Subjektive Maße zur Erfassung der Nutzerzustandsdimensionen (Experiment 4)

Nutzerzustand	Frage/Methode
Beanspruchung	• Skala Anstrengung des NASA-TLX: "Wie sehr mussten Sie sich anstrengen, um die gestellten Aufgaben zu erfüllen?"
Aufmerksamkeit	• Selbst generierte Frage: „Wie bewerten Sie den Grad Ihrer Aufmerksamkeit (Wahrnehmung wichtiger Ereignisse und Aufgaben)?"
Müdigkeit	• Selbst generierte Frage: „Wie bewerten Sie den Grad Ihrer Müdigkeit?"
Motivation	• Selbst generierte Frage: „Wie hoch ist Ihre Motivation, bei der Untersuchung eine gute Leistung zu erzielen?"
Emotionaler Zustand	• Bewertung hinsichtlich der drei Dimensionen Valenz, Arousal und Dominanz des SAM (Bradley & Lang, 1994)
Level-1-SA	• „Welche Aufgaben mussten zum Zeitpunkt der Unterbrechung bearbeitet werden? Antwortmöglichkeiten: „Identifizieren außerhalb der ISR", „Identifizieren innerhalb der ISR", „Warnen", „Bekämpfen", „NRTT-Kontakt anlegen", „Keine Aufgabe" (Mehrfachauswahl möglich)

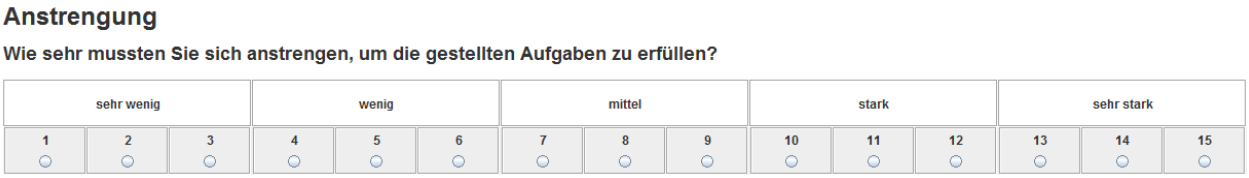


Abbildung 46. Darstellung der in Experiment 4 verwendeten KU-Skala (Heller, 1982)

Um die Anzahl an Fragen zu reduzieren, wurde die *mentale Beanspruchung* eindimensional erfasst. Analog zu der Ratingskala *RSME* (Rating Scale of Mental Effort, Zijlstra, 1993) wurde hierfür die Dimension *Anstrengung* herangezogen, die auch eine der Subskalen des NASA-TLX darstellt.

Für *Müdigkeit* sind validierte Fragebögen oder Skalen, wie die *Stanford Sleepiness Scale* (Hoddes et al., 1973) oder die *Karolinska Sleepiness Scale* (Åkerstedt & Gillberg, 1990) verfügbar (vgl. ausführliche Übersicht in Gawron, 2016). Diese sind jedoch meist darauf ausgelegt, die schlafbezogene Müdigkeit zu erfassen und daher für die Erfassung der aufgabenbezogenen Müdigkeit weniger geeignet; oder sie bestehen aus mehreren Items/Fragen und sind daher zu zeitintensiv. Müdigkeit wurde daher, ebenso wie die *Aufmerksamkeit* und die *Motivation*, über eine selbst generierte Frage erfasst. Hinsichtlich der Aufmerksamkeit ist jedoch, wie bereits in Experiment 1 angemerkt wurde (vgl. Abschnitt 4.2.6), zu bedenken, dass eine subjektive Bewertung der Aufmerksamkeit fehlerbehaftet sein könnte. Zur Validierung kritischer Aufmerksamkeit wurde daher zusätzlich das *Level-1-SA* als Vergleichsmaß herangezogen.

Level-1-SA bezieht sich auf die Wahrnehmung wichtiger Elemente und Ereignisse in der gegenwärtigen Situation und ist damit eng mit der Aufmerksamkeitsverteilung verbunden (vgl. Abschnitt 3.1.6). Level-1-SA wurde in Anlehnung an das Verfahren *SAGAT* (Situation Awareness Global Assessment Technique, Endsley, 1988) indirekt erfasst. Nach Unterbrechung des Szenarios bei einem Leistungseinbruch sollte der Proband – ohne das Lagebild sehen zu können – angeben, welche Aufgaben zum Zeitpunkt der Unterbrechung bearbeitet werden mussten (vgl. Tabelle 47). Um den Einfluss von Gedächtniseffekten zu reduzieren, wurde diese Frage direkt zu Beginn der Befragung gestellt. Zur Bewertung von Level-1-SA wurden die vom Probanden angegebenen Aufgaben mit den Aufgaben verglichen, die tatsächlich bearbeitet werden mussten. Wenn die Aufgabe, die den Leistungseinbruch ausgelöst hatte, vom Probanden nicht als zu bearbeitende Aufgabe genannt worden war, wurde dies als fehlerhaftes Level-1-SA gewertet. Ein Vorteil gegenüber der subjektiven Bewertung besteht darin, dass Fehleinschätzungen vermieden werden, die entstehen können, wenn sich die Teilnehmer nicht bewusst sind, dass ihr Situationsbewusstsein fehlerhaft ist. Aufgrund dieser Eigenschaft, und da korrektes Level-1-SA eine korrekte Aufmerksamkeitsverteilung voraussetzt, wird dieses Maß in der Validierungsstudie als zusätzliches Referenzmaß für die Aufmerksamkeit herangezogen.

7.1.5 Hypothesen

Zur Validierung der Diagnoseergebnisse von RASMUS wurde untersucht, ob die von RASMUS diagnostizierten kritischen Nutzerzustände zum Zeitpunkt von Leistungseinbrüchen mit den subjektiven Bewertungen des Nutzerzustands korrespondieren. Zum Einen wird angenommen, dass das subjektive Vergleichsmaß bei Leistungseinbrüchen, bei denen der betrachtete Nutzerzustand gemäß der RASMUS-Diagnose kritisch ausgeprägt ist, in erwarteter Richtung von der Baseline abweicht. Bei kritischer Beanspruchung und Müdigkeit wird eine positive, bei kritischer Aufmerksamkeit eine negative Abweichung erwartet. Zum Anderen wird angenommen, dass das Vergleichsmaß bei kritischer Ausprägung eine stärkere Abweichung aufweist als bei unkritischer Ausprägung des Nutzerzustands.

In Hinblick auf die Aufmerksamkeit wird außerdem angenommen, dass das Level-1-SA bei Leistungseinbrüchen mit kritischer Aufmerksamkeit häufiger fehlerhaft ist als bei den übrigen Leistungseinbrüchen ohne kritische Aufmerksamkeit. In Tabelle 48 sind die Hypothesen für die diagnostizierten kritischen Nutzerzustände zusammengefasst. Die Annahmen zu den Abweichungen von der Baseline wurden deskriptiv ausgewertet. Die Unterschiede zwischen kritischen und unkritischen Nutzerzuständen wurden für jedes Vergleichsmaß inferenzstatistisch geprüft (vgl. Abschnitt 7.1.8).

Tabelle 48. Hypothesen zur Validierung der von RASMUS diagnostizierten Nutzerzustände bei Leistungseinbrüchen (Experiment 4)

Hypothese	Diagnostizierter Nutzerzustand	Vergleichsmaß	Erwartete Abweichung von Baseline	Erwarteter Unterschied zum unkritischen Zustand
H1	Kritische Beanspruchung	NASA-TLX Skala Anstrengung	Positive Abweichung	$\mu_{\text{krit. B.}} > \mu_{\text{nicht krit. B.}}$
H2	Kritische Müdigkeit	Subjektive Müdigkeit	Positive Abweichung	$\mu_{\text{krit. M.}} > \mu_{\text{nicht krit. M.}}$
H3a	Kritische Aufmerksamkeit	Subjektive Aufmerksamkeit	Negative Abweichung	$\mu_{\text{krit. A.}} < \mu_{\text{nicht krit. A.}}$
H3b		Fehlerhaftes Level-1-SA	Keine Baseline vorhanden	$\pi_{\text{krit A.}} > \pi_{\text{nicht krit. A.}}$

Legende: krit. B. = kritische Beanspruchung, krit. M.= kritische Müdigkeit, krit A.= kritische Aufmerksamkeit, LE = Leistungseinbruch, $\pi_{\text{krit A.}}$ = Häufigkeit fehlerhaftes SA bei LE mit kritischer Aufmerksamkeit, $\pi_{\text{nicht krit,A.}}$ = Häufigkeit fehlerhaftes SA bei LE mit nicht kritischer Aufmerksamkeit.

7.1.6 Versuchsdurchführung

Vor Durchführung des Experiments wurde eine Pilotstudie mit drei Teilnehmern durchgeführt, um die korrekte Funktionsweise der Echtzeitdiagnose zu überprüfen, und zu testen, ob das Szenario geeignet ist, Leistungseinbrüche und kritische Nutzerzustände hervorzurufen. Dabei zeigte sich, dass eine Anpassung der Regel für kritische Aufmerksamkeit notwendig ist (vgl. Abschnitt 6.6.2). Außerdem wurden kleinere Fehler bei der Implementierung identifiziert und vor Beginn des Validierungsexperiments behoben.

Zu Beginn des Versuchs wurden dem Versuchsteilnehmer allgemeine Informationen zum Versuch und eine Einverständniserklärung in schriftlicher Form vorgelegt (s. Anhang E.1 und A.2), die er/sie durch Unterschrift bestätigen sollte. Danach folgte das Ausfüllen eines kurzen Fragebogens in elektronischer Form, in dem Alter, Geschlecht, Erfahrung sowie das aktuelle Befinden (Müdigkeit, emotionaler Zustand, Motivation) erfragt wurden. Anschließend erfolgten das Anlegen des BioHarness sowie das Kalibrieren des Eyetrackers. Die Durchführung der Experimentalaufgaben wurde dem Versuchsteilnehmer schriftlich (vgl. Anhang E.1) und auch mündlich im Rahmen eines Übungsszenarios erläutert. Im Übungsszenario konnte der Proband die Aufgaben mehrere Male selbst ausführen und bei Unklarheiten Fragen stellen. Am Ende des Übungsszenarios – sowie auch am Ende des späteren Versuchsszenarios – wurde der Teilnehmer gebeten, seinen aktuellen Nutzerzustand zur Erfassung der Baseline in einem Fragebogen zu bewerten.

Nach erfolgreichem Absolvieren des Übungsszenarios wurde das Versuchsszenario gestartet. Die Dauer des Szenarios betrug 45 Minuten, wobei sich die tatsächliche Dauer durch das Pausieren und das Beantworten des Fragebogens bei Leistungseinbrüchen verlängerte. Insgesamt lag die Dauer des Versuchs bei 1,5 bis 2 Stunden. Start- und Endzeitpunkt sowie wichtige Beobachtungen während des Versuchs wurden vom Versuchsleiter in einem Protokoll (siehe Anhang E.2) notiert.

7.1.7 Stichprobe

Das Experiment wurde mit 12 Versuchsteilnehmern durchgeführt. Dabei handelte es sich um 9 männliche und 3 weibliche Mitarbeiter des FKIE im Alter zwischen 24 und 43 Jahren (M=33). 5 Teilnehmer gaben gute bis sehr gute Kenntnisse mit computerbasierten Simulations- und Strategiespielen an (alle männlich), die anderen 7 Teilnehmer geringe bis gar keine Kenntnisse. Die Teilnehmer hatten keine praktischen Erfahrungen mit der Bearbeitung von Aufgaben aus dem Bereich AAW. Einigen Teilnehmern war die Benutzungsoberfläche allerdings aus vorherigen Projekten bekannt.

7.1.8 Datenaufbereitung und -auswertung

Im Rahmen der Datenaufbereitung wurden die Ergebnisse der subjektiven Befragung anhand der Zeitstempel den Diagnoseergebnissen von RASMUS bei Leistungseinbrüchen zugeordnet. Die Auswertung erfolgte mit dem Statistikprogramm SPSS 20.0. Da es bei den verschiedenen Versuchsteilnehmern unterschiedlich oft zu Leistungseinbrüchen kam, und die verschiedenen kritischen Nutzerzustände unterschiedlich häufig pro Person identifiziert wurden, setzt sich der Datensatz aus unterschiedlich vielen messwiederholten („within subjects“-) Daten und Daten zwischen Personen („between-subjects“) zusammen. Üblicherweise werden diese Daten nicht gemischt ausgewertet, da es zu einer Konfundierung zwischen individuellen Einflüssen und Effekten der interessierenden unabhängigen Variablen kommen kann (Bland & Altman, 1994). Eine ausschließliche Analyse der Daten zwischen Personen oder innerhalb von Personen würde andererseits bedeuten, dass ein Großteil der Daten nicht berücksichtigt werden kann, was die Aussagekraft der Ergebnisse ebenfalls einschränken würde. Für diesen Fall ist die Anwendung von Multilevel-Modellen bzw. gemischten Modellen angebracht, bei denen die Personen als zufälliger Faktor einbezogen werden. Diese Verfahren benötigen jedoch für eine zuverlässige Schätzung der Parameter große Stichprobenumfänge pro Analyseebene (Tabachnick & Fidell, 2007, S. 787 ff.). In Anbetracht der Fallzahlen im Validierungsexperiment, die in Abschnitt 7.2.1 und in Anhang E.3 Abbildung 56 auf Personenebene aufgeführt sind, scheint dies nicht gegeben zu sein. Aus diesem Grund wird es als zweckmäßiger erachtet, die Personenzugehörigkeit nicht in die inferenz-statistische Analyse miteinzubeziehen und die Effekte zusätzlich auf Individualebene deskriptiv zu analysieren. Diese Analyse ist ebenfalls in Anhang E.3 in Abbildung 57 aufgeführt.

Die Hypothesen H1 bis H3a wurden daher über ungepaarte t-Tests untersucht. Nach Bortz (2005) wird die Präzision eines t-tests bei unterschiedlichen Stichprobenumfängen der zu vergleichenden Gruppen (in diesem Fall die unterschiedliche Anzahl an kritischen und unkritischen Zuständen) nicht beeinträchtigt, wenn die Varianzen gleich sind. Neben der Normalverteilung der Daten wurde daher die Varianzhomogenität der Gruppen über den Levene-Test geprüft. Bei nicht gegebenen

Voraussetzungen wurde der nonparametrische Mann-Whitney-U-Test verwendet Bezüglich Hypothese H3b wurde der χ^2-Test angewendet, um zu überprüfen, ob die Häufigkeit von fehlerhaftem Level-1-SA bei Leistungseinbrüchen mit kritischer und unkritischer Aufmerksamkeit statistisch signifikant ausfällt.

7.2 Ergebnisse

Der Ergebnisteil gliedert sich in eine deskriptive Analyse der Daten (Abschnitt 7.2.1), eine inferenzstatistische Untersuchung auf Basis der in Abschnitt 7.1.5 definierten Hypothesen (Abschnitt 7.2.2) sowie eine Post hoc-Analyse zur Güte der diagnostischen Entscheidungen (Abschnitt 7.2.3).

7.2.1 Deskriptive Analyse

Dieser Abschnitt gibt zunächst einen Überblick über die prozentualen Häufigkeiten diagnostizierter Nutzerzustände und Leistungseinbrüche in den Phasen des Versuchsszenarios. Außerdem stellt er dar, wie häufig die jeweiligen Nutzerzustände bei Leistungseinbrüchen kritisch ausgeprägt waren.

Analyse hoher Beanspruchung und passiver aufgabenbezogener Müdigkeit pro Phase

Wie in Abschnitt 7.1.3 beschrieben, setzt sich das Versuchsszenario aus den drei Phasen *Baseline*, *hohe Belastung* und *Monotonie* zusammen, wobei in der Phase hoher Belastung insbesondere Zustände hoher Beanspruchung hervorgerufen werden sollten, während in der Monotoniephase passive aufgabenbezogene Müdigkeit erzeugt werden sollte. In der Baseline-Phase sollte keiner dieser Zustände provoziert werden. Um zu überprüfen, ob dies gelungen ist, wurde zunächst ermittelt, wie häufig hohe Beanspruchung und passive aufgabenbezogene Müdigkeit den Diagnoseregeln zufolge in den einzelnen Phasen des Versuchs aufgetreten sind[7]. Der durchschnittliche Anteil der Zeit, in dem diese beiden Zustände von RASMUS diagnostiziert wurden, ist pro Phase in Abbildung 47 wiedergegeben.

Es zeigt sich, dass in der Baseline-Phase nur zu einem sehr geringen Anteil hohe Beanspruchung und passive aufgabenbezogene Müdigkeit diagnostiziert wurden: Bei hoher Beanspruchung liegt der Anteil bei 3,2% und bei Müdigkeit bei 10,2%. Erwartungsgemäß ist der Prozentanteil hoher Beanspruchung in der Phase hoher Belastung sehr viel höher (37,3%) als in der Baseline-Phase und der Monotonie-Phase. Außerdem ist der Prozentanteil passiver aufgaben-bezogener Müdigkeit ebenfalls erwartungsgemäß in der Monotonie-Phase deutlich höher (50,7%) als in der Baselinephase und der Phase hoher Belastung. Die Ergebnisse weisen somit darauf hin, dass die beiden Nutzerzustände *hohe Beanspruchung* und *passive aufgabenbezogene Müdigkeit* erfolgreich durch die Phasen moduliert werden konnten.

[7] Diese Analyse erfolgte über eine Videoauswertung der im Leistungs- und Zustandsmonitor dargestellten kritischen Nutzerzustände, da die Diagnoseergebnisse von RASMUS entsprechend dem Diagnosekonzept nur zum Zeitpunkt von Leistungseinbrüchen mitgeloggt werden (vgl. Abschnitt 6.2.4).

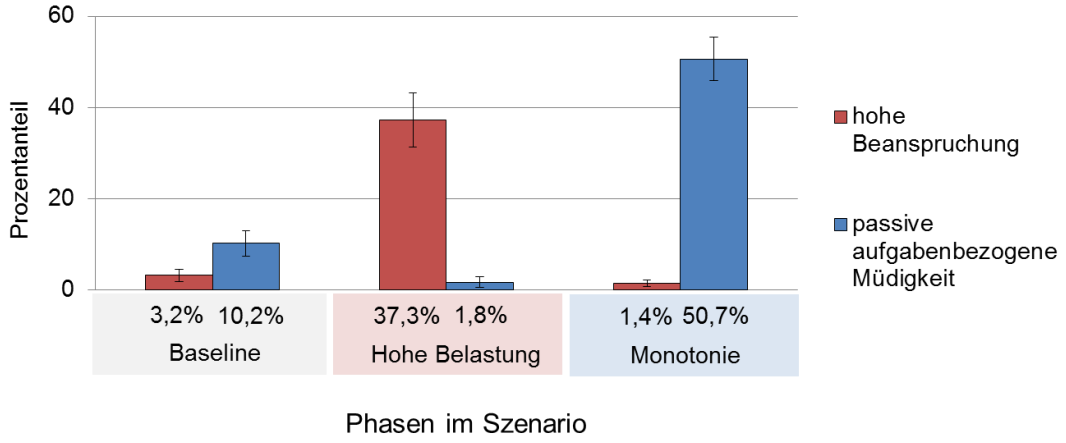


Abbildung 47. Durchschnittliche Zeitanteile hoher Beanspruchung und passiver aufgabenbezogener Müdigkeit pro Phase in Prozent (Fehlerbalken: Standardfehler) – Experiment 4

Analyse von Leistungseinbrüchen

In die deskriptive und inferenzstatistische Analyse wurden nur die Leistungseinbrüche einbezogen, die eine Unterbrechung des Szenarios und eine Befragung zum Nutzerzustand auslösten. Leistungseinbrüche, die in der „Totzeit“ (vgl. Abschnitt 6.2.2), d.h. innerhalb von 60 Sekunden nach dem letzten Leistungseinbruch, auftraten, wurden nicht betrachtet. Insgesamt sind 76 Leistungseinbrüche aufgetreten, die für die Validierung der Echtzeitdiagnose herangezogen werden konnten. In Abbildung 48 ist die Anzahl an Leistungseinbrüchen pro Minute des Szenarios über alle Probanden hinweg in einem Balkendiagramm abgetragen. Die Farbunterteilung innerhalb der Balken gibt dabei Aufschluss über die Häufigkeit der als kritisch diagnostizierten Nutzerzustände.

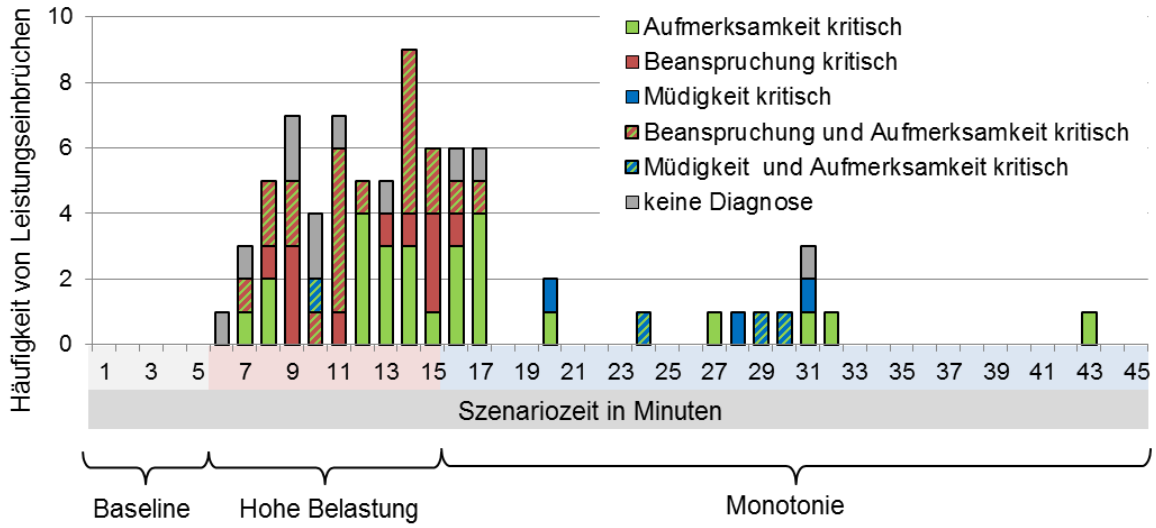


Abbildung 48. Verteilung von Leistungseinbrüchen im Szenarioverlauf (Experiment 4)

Es wird ersichtlich, dass in der Baseline-Phase, wie vorgesehen, kein Leistungseinbruch aufgetreten ist. Die meisten Leistungseinbrüche haben sich erwartungsgemäß in der Phase hoher Belastung ereignet, in der die Nutzerzustände Beanspruchung und Aufmerksamkeit überwiegend als kritisch diagnostiziert wurden. Einige Leistungseinbrüche mit hoher Beanspruchung fallen durch den fließenden Übergang von hoher Belastung zu Monotonie allerdings auch noch in den Beginn der

Monotonie-Phase. Nach der Übergangsphase traten in der Monotonie-Phase nur sehr vereinzelt Leistungseinbrüche auf, die erwartungskonform mit kritischer Müdigkeit und/oder kritischer Aufmerksamkeit verbunden waren.

Die Häufigkeit von Leistungseinbrüchen pro Person reicht von 1 bis 14 und liegt im Durchschnitt bei *M*=6,3 (*SD*=3,6). Dies weist auf starke interindividuelle Unterschiede in der Leistung hin. Ein Einflussfaktor könnte die Erfahrung im Umgang mit computerbasierten Simulations- und Strategiespielen sein. So weisen Personen mit guten Kenntnissen im Durchschnitt weniger Leistungseinbrüche auf (*M*=5,2; *n*=5) als Personen mit geringen Kenntnissen (*M*=7,1; *n*=7). Dieser Unterschied erwies sich in einem Mann-Whitney-U-Test allerdings als nicht signifikant (*p*=.36).

Analyse kritischer Nutzerzustände bei Leistungseinbrüchen

Tabelle 49 gibt die Fallzahlen für die Diagnoseergebnisse von RASMUS bei Leistungseinbrüchen wieder. In Hinblick auf kritische Aufmerksamkeit wird zusätzlich unterschieden, ob diese diagnostiziert wurde, weil nur eine Aufgabe vorlag, die nicht bearbeitet wurde, oder weil bei Vorhandensein von mehreren Aufgaben eine Aufgabe mit geringerer Priorität bearbeitet wurde.

Tabelle 49. Häufigkeiten diagnostizierter kritischer und unkritischer Nutzerzustände bei Leistungseinbrüchen in Experiment 4

Aufmerksamkeit	Beanspruchung kritisch	Müdigkeit kritisch	Beanspruchung und Müdigkeit unkritisch	Summe
Kritisch – Aufgabe nicht bearbeitet	2	3	6	**11**
Kritisch – Priorität zu niedrig	19	1	20	**40**
Unkritisch	11	3	11	**25**
Summe	**32**	**7**	**37**	**76**

Aus Tabelle 49 geht hervor, dass bei Leistungseinbrüchen am häufigsten eine falsche Priorisierung der bestehenden Aufgaben vorlag (*n*=40). Die falsche Priorisierung fiel in etwa der Hälfte der Fälle mit kritischer Beanspruchung zusammen, nur in einem Fall mit kritischer Müdigkeit. Letzterer Befund ist damit zu erklären, dass Müdigkeit, wie aus Abbildung 48 hervorgeht, hauptsächlich in der Monotoniephase kritisch ausgeprägt war, in der nur selten mehr als eine Aufgabe zur gleichen Zeit bearbeitet werden musste und damit nur selten die Voraussetzung für eine falsche Priorisierung gegeben war.

Es zeigt sich außerdem, dass mehr Leistungseinbrüche mit kritischer Beanspruchung (*n*=32) als mit kritischer Müdigkeit (*n*=7) verbunden sind. Dies war ebenfalls zu erwarten, da in der Phase hoher Belastung, in der die Beanspruchung am häufigsten kritisch ausgeprägt war, mehr Aufgaben bearbeitet werden mussten und somit mehr Leistungseinbrüche auftreten konnten als in der Müdigkeit erzeugenden Monotoniephase. Bei 11 Leistungseinbrüchen wurde kein kritischer Nutzerzustand diagnostiziert. Mögliche Gründe dafür werden in Abschnitt 7.3.4 diskutiert.

Analyse auf Individualebene

In Anhang E.3 Abbildung 56 sind die Häufigkeiten kritischer und unkritischer Nutzerzustände bei Leistungseinbrüchen für jede Versuchsperson separat aufgeführt. Es wird ersichtlich, dass sich nicht nur die Häufigkeit der Leistungseinbrüche sondern auch die Häufigkeit der jeweiligen kritischen Nutzerzustände bei Leistungseinbrüchen zwischen den Personen stark unterscheidet. In Bezug auf Müdigkeit wurde nur bei vier Personen ein kritischer Zustand diagnostiziert. Dies hat zur Folge, dass die Versuchspersonen die nachfolgenden Untersuchungsergebnisse unterschiedlich stark beeinflusst haben. Wie bereits in Abschnitt 7.1.8 erläutert wurde, werden die Ergebnisse daher auch auf Individualebene betrachtet. Die Analysen sind in Anhang E.3 Abbildung 57 aufgeführt und werden in die Diskussion der Ergebnisse in Abschnitt 7.3 einbezogen.

7.2.2 Hypothesenprüfende Untersuchung

In Abbildung 49 sind die Ergebnisse zu den Abweichungen der subjektiven Vergleichsmaße von der Baseline als Balkendiagramm dargestellt. Es zeigt sich, dass die subjektiven Vergleichsmaße bei Leistungseinbrüchen mit dem Diagnoseergebnis *kritische Beanspruchung* und *kritische Müdigkeit* erwartungsgemäß positiv von der Baseline abweichen. Die Aufmerksamkeit wird jedoch entgegen der Annahme bei Leistungseinbrüchen mit dem Diagnoseergebnis *kritische Aufmerksamkeit* nicht geringer beurteilt als bei der Baseline. In Abbildung 49 wird auch ersichtlich, dass Beanspruchung und Müdigkeit bei kritischer Ausprägung dieser Zustände entsprechend der RASMUS-Diagnose subjektiv jeweils höher bewertet werden als bei unkritischer Ausprägung. Die Aufmerksamkeit wird bei kritischer Ausprägung hingegen nur marginal geringer bewertet als bei unkritischer Ausprägung.

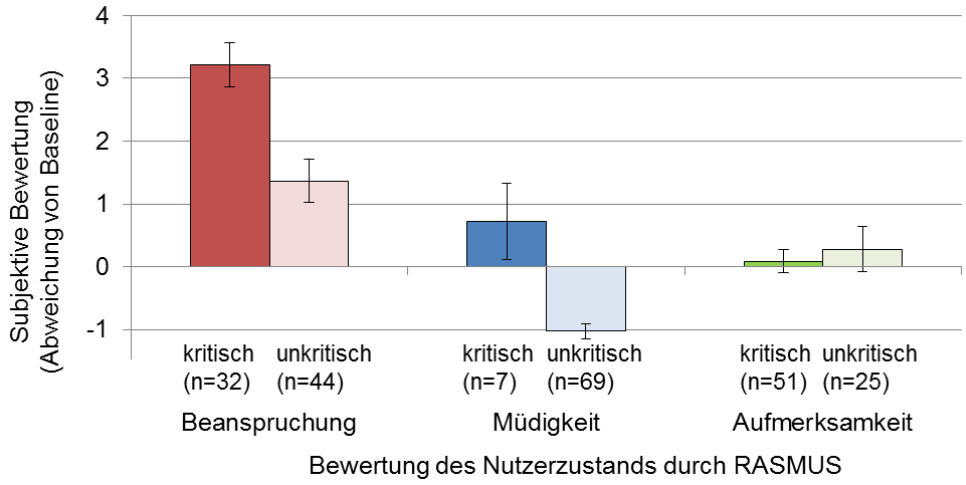


Abbildung 49. Durchschnittliche Abweichungen der subjektiven Bewertungen von der Baseline für kritische und unkritische Ausprägungen der Nutzerzustände nach RASMUS (Fehlerbalken: Standardfehler) – Experiment 4

Die Unterschiede zwischen kritischen und unkritischen Ausprägungen wurden über t-Tests auf Signifikanz getestet. Bei Prüfung der Voraussetzungen konnte die Varianzhomogenität für alle Variablen bestätigt werden. Allerdings ist die Annahme der Normalverteilung der Daten nicht für alle Variablen gegeben, so dass zusätzlich der nonparametrische Mann-Whitney-U-Test berechnet wurde.

Wie sich in Tabelle 50 zeigt, fallen die Ergebnisse des Mann-Whitney-U-Tests identisch mit denen des t-Tests aus. In beiden Analysen erweisen sich für die Nutzerzustände Beanspruchung und Müdigkeit die Unterschiede in der subjektiven Bewertung bei kritischer und unkritischer Ausprägung als signifikant. Für Aufmerksamkeit liegt hingegen kein signifikanter Unterschied in der subjektiven Bewertung vor. Die Analysen sprechen somit dafür, dass die Hypothesen 1 und 2 bestätigt werden können, wohingegen Hypothese 3a nicht bestätigt werden kann.

Tabelle 50. Ergebnisse der inferenzstatistischen Auswertung zur Überprüfung der Hypothesen H1-H3a (Experiment 4)

	H1: Beanspruchung		**H2: Müdigkeit**		**H3a: Aufmerksamkeit**	
	t(74)	*p*	*t*(74)	*p*	*t*(74)	*p*
t-Test	3.67	<.001**	3.99	<.001**	-.52	.60
Mann-Whitney U-Test	*z* = -3.5	<.001**	*z* = -2.6	.009**	*z* = -.05	.96

Anmerkungen: AV: subjektives Vergleichsmaß (Abweichung von der Baseline); ** p <.01

Um Hypothese H3b zu prüfen, werden im Folgenden die Unterschiede im Level-1-SA zwischen kritischer und unkritischer Aufmerksamkeit entsprechend der RASMUS-Diagnose analysiert. Abbildung 50 gibt die Häufigkeiten von korrektem und falschem Level-1-SA für Leistungseinbrüche mit kritischer und mit nicht kritischer Aufmerksamkeit wieder. Dabei zeigt sich hypothesenkonform, dass falsches SA bei Leistungseinbrüchen mit kritischer Aufmerksamkeit häufiger vorkommt als bei nicht kritischer Aufmerksamkeit. Für die statistische Testung wurde der χ^2-Test angewandt. Dabei erwiesen sich die Häufigkeitsunterschiede zwischen kritischer und nicht kritischer Aufmerksamkeit als signifikant (χ^2 (1) = 15.54, p<.001). Hypothese H3b kann somit bestätigt werden.

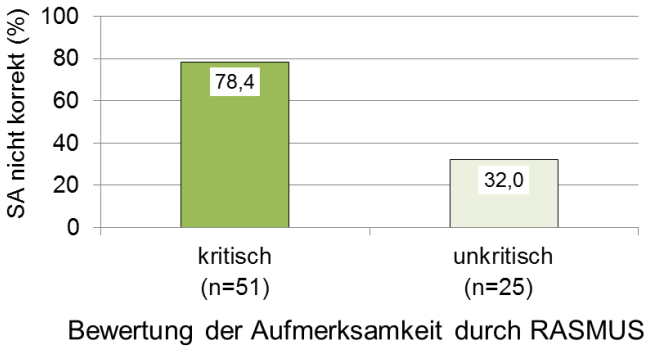


Abbildung 50. Prozentuale Häufigkeit von nicht korrektem Level-1-SA bei Leistungseinbrüchen (LE) mit kritisch und nicht kritisch bewerteter Aufmerksamkeit durch RASMUS (Experiment 4)

7.2.3 Analyse zur Güte der diagnostischen Entscheidung

Bei der hypothesenprüfenden Untersuchung wurde (mit Ausnahme der Analyse zu H3b) untersucht, ob die subjektiven Vergleichsmaße bei Leistungseinbrüchen in erwarteter Richtung von der Baseline abweichen und sich signifikant zwischen den Diagnoseergebnissen von RASMUS unterscheiden. Dabei ist zu bedenken, dass starke Abweichungen von der Baseline bei einzelnen Leistungseinbrüchen gegenläufige Abweichungen bei anderen Leistungseinbrüchen überdecken

können. Die Ergebnisse geben somit keinen Aufschluss darüber, wie oft die Diagnoseergebnisse von RASMUS im Einzelfall mit den subjektiven Vergleichsmaßen übereinstimmen. Die Übereinstimmung der RASMUS-Diagnose mit der subjektiven Bewertung soll daher auch anhand von diagnostischen Kennwerten bewertet werden, die sich auf die Häufigkeit richtiger und falscher Diagnoseentscheidungen beziehen.

Vorgehen

Die Beurteilung, ob die Diagnose eines Nutzerzustands als *kritisch* bzw. *unkritisch* richtig oder falsch war, setzt Kenntnis darüber voraus, ob der wahre zu diagnostizierende Zustand kritisch oder unkritisch ausgeprägt war. Da im vorliegenden Fall der wahre Nutzerzustand nicht bekannt ist, wird erneut die subjektive Bewertung des Nutzerzustands als Vergleichsmaß herangezogen. Um die diagnostischen Kennzahlen zu ermitteln, muss diese zunächst von einer metrischen in eine dichotome Variable (Zustand kritisch vs. unkritisch) transformiert werden. Als Trennwert für die Unterscheidung zwischen einem kritischen und einem unkritischen Zustand wurde der individuelle Baseline-Wert herangezogen. Das heißt, die Diagnose eines kritischen Nutzerzustands durch RASMUS wird dann als korrekt bewertet, wenn auch die subjektive Bewertung des Nutzerzustands in erwarteter Richtung von der Baseline abweicht (so genannte *richtig positive* Klassifizierung). Demgegenüber wird die Diagnose eines kritischen Zustands als falsch bewertet, wenn die subjektive Bewertung entweder nicht abweicht oder eine anders gerichtete Abweichung aufweist (sog. *falsch positive* Klassifizierung).

Analog wird auch für die Diagnose eines unkritischen Zustands die Übereinstimmung mit der subjektiven Bewertung ermittelt. Bei Übereinstimmung (d.h. es liegt keine oder eine anders gerichtete Abweichung von der Baseline vor) gilt die Klassifizierung als *richtig negativ*. Bei Nichtübereinstimmung wird sie als *falsch negativ* gewertet. Anhand der relativen Häufigkeiten dieser vier Befunde können verschiedene Kennwerte berechnet werden, die Aufschluss über die Güte der diagnostischen Entscheidungen geben sollen. Dazu zählen die *Sensitivität* und *Spezifität*, sowie der *positive* und *negative prädiktive Wert* (s. Glossar für eine Definition).

Ergebnisse

Tabelle 51 gibt die Häufigkeiten richtig/falsch positiver und richtig/falsch negativer Klassifizierungen für die drei Nutzerzustände wieder. Zur besseren Unterscheidung ist die Anzahl der richtig positiven und richtig negativen Klassifizierungen grün und die Anzahl der falsch positiven und falsch negativen Klassifizierungen rot dargestellt. In den Zeilen darunter sind zu jedem Nutzerzustand die sich daraus ergebenden diagnostischen Kennwerte aufgeführt.

Tabelle 51. Diagnostische Kennwerte zur Güte der Diagnosefähigkeiten von RASMUS (bei Verwendung der dichotomisierten subjektiven Vergleichsmaße als Referenz) – Experiment 4

		Ergebnis Subjektive Bewertung							
		Beanspruchung		Müdigkeit		Aufmerksamkeit			
		> BL	<= BL	> BL	<= BL	< BL	>= BL	SA inkorrekt	SA korrekt
Ergebnis RASMUS	kritisch	28[a]	4[b]	5[a]	2[b]	17[a]	34[b]	40[a]	11[b]
	unkritisch	30[c]	14[d]	7[c]	62[d]	10[c]	15[d]	8[c]	17[d]
Positiver prädiktiver Wert		87,5%		71,4%		33,3%		78,4%	
Negativer prädiktiver Wert		32%		89,9%		60%		68%	
Sensitivität		48,3%		41,7%		63%		83,3%	
Spezifität		77,8%		96,9%		30,6%		60,7%	

BL = Baseline, [a] richtig Positive, [b] falsch Positive, [c] falsch Negative, [d] richtig Negative

Interpretation der Ergebnisse

Aus den Angaben zu den positiven prädiktiven Werten in Tabelle 51geht hervor, dass die Mehrheit der Leistungseinbrüche, bei denen von RASMUS eine kritisch hohe Beanspruchung bzw. Müdigkeit diagnostiziert wurde, auch mit einer höheren subjektiven Bewertung der Beanspruchung bzw. Müdigkeit einhergeht. In Bezug auf Aufmerksamkeit ist dies nicht der Fall, wenn die subjektiv bewertete Aufmerksamkeit als Vergleichsmaß herangezogen wird. Wird jedoch das Level-1-SA als Kriterium für richtige und falsche Diagnosen der Aufmerksamkeitsverteilung herangezogen, liegt der positive prädiktive Wert bei 78,4%.

In Hinblick auf Leistungseinbrüche, bei denen der jeweilige Nutzerzustand von RASMUS als unkritisch bewertet wurde, stimmt das Diagnoseergebnis bei Müdigkeit und Aufmerksamkeit ebenfalls mehrheitlich mit der subjektiven Bewertung überein (vgl. Ergebnisse zu den negativen prädiktiven Werten in Tabelle 51). Die Beanspruchung wurde subjektiv hingegen bei Leistungseinbrüchen mit unkritischer Beanspruchung häufiger höher bewertet als bei der Baseline.

Zu beachten ist, dass die prädiktiven Werte von der Prävalenz des interessierenden Merkmals (in diesem Fall dem kritischen Nutzerzustand) beeinflusst werden. In Hinblick auf die Beanspruchung liegt bei der Mehrheit der Leistungseinbrüche ein kritischer Zustand vor, wenn die Abweichung des subjektiven Vergleichsmaßes von der Baseline als Referenz verwendet wird. Dies führt zu einer hohen Fallzahl bei den richtig positiven und falsch negativen Klassifizierungen und trägt damit zu einem im Vergleich höheren positiven und geringeren negativen prädiktiven Wert bei.

Ähnlich verhält es sich mit den Kennwerten Sensitivität und Spezifität. Sie sind zwar unabhängig von der Prävalenz, allerdings werden sie von der Auftretenshäufigkeit kritischer und unkritischer Diagnoseergebnisse beeinflusst. Aus Tabelle 51 geht hervor, dass die Nutzerzustände Beanspruchung und Müdigkeit nach RASMUS seltener kritisch als unkritisch ausgeprägt sind. Dies trägt dazu bei, dass die Spezifität hohe Werte aufweist und bei Müdigkeit nahe 100% liegt, während die Sensitivität deutlich schwächer ausfällt. Bei der Aufmerksamkeit sind die Befunde umgekehrt. Auffällig ist, dass bei Verwendung von Level-1-SA als Vergleichsmaß sowohl Sensitivität als auch Spezifität deutlich höher ausfallen als bei der subjektiven Aufmerksamkeit.

Bewertung über die Receiver Operating Characteristic (ROC)-Curve

Wie sich im vorigen Abschnitt zeigte, ist die Betrachtung eines einzelnen diagnostischen Kennwerts für die Bewertung der Güte eines Diagnosemaßes nicht ausreichend. Im Extremfall weist ein Diagnosemaß, das jeden Zustand als kritisch bewertet, eine Sensitivität von 100%, dafür aber nur eine sehr geringe Spezifität auf. Verschiedene Verfahren kombinieren daher diese Kennwerte, um allgemeine Aussagen zur diagnostischen Güte zu treffen, z.B. *Youden*-Index (Youden, 1950), *Likelihood-Ratio* (LR; Deeks & Altman, 2004), *Diagnostic Odds Ratio* (DOR; Glas et al., 2003) und die *Receiver Operating Characteristic (ROC-) Curves* mit dem Index *Area Under Curve* (AUC; vgl. Harley & McNeil, 1982; Zweig & Campbell, 1994; Swets, 1996).

Im Folgenden soll eine allgemeine Bewertung der Diagnoseergebnisse von RASMUS mit Hilfe der ROC-Kurven vorgenommen werden (s. Glossar für eine Erläuterung). Zur Bewertung der Diagnosegüte wird die Fläche unterhalb der ROC-Kurve, die sogenannte *Area under Curve (AUC)* herangezogen. Diese liegt bei einer Klassifizierung nach dem Zufallsprinzip bei 0.5 (die ROC-Kurve entspricht dann einer Diagonalen) und bei einer perfekten Klassifizierung bei 1. Da die Berechnung der AUC dem Wilcoxon Signed-Rank-Verfahren gleicht, kann außerdem geprüft werden, ob die Fläche signifikant von 0.5 - und damit dem Ergebnis bei einer zufälligen Klassifizierung – abweicht (Zweig & Campbell, 1994; Hanley & McNeil, 1982).

Das Verfahren ermöglicht es auch, die Diagnosegüte für verschiedene Trennwerte eines Klassifikators zu bewerten. Aus Tabelle 51 geht hervor, dass die Verwendung der Baseline als Trennwert zwischen kritischen und unkritischen Zuständen bei Beanspruchung und Müdigkeit zu einer im Vergleich zur RASMUS-Diagnose hohen Zahl kritischer Zustände führt. Da dieser Trennwert frei gewählt wurde, und geringfügige Abweichungen von der Baseline möglicherweise noch keinen kritischen Zustand anzeigen, soll mittels der ROC-Kurven auch geprüft werden, ob bei Veränderung des Trennwerts eine bessere Übereinstimmung mit der RASMUS-Diagnose erzielt werden kann.

In Tabelle 52 sind die Ergebnisse dieser Analysen aufgeführt. Bei Beanspruchung ergeben sich durch Erhöhung des Trennwerts auf 3 Skalenwerte über der Baseline deutlich bessere diagnostische Kennwerte als bei Verwendung der Baseline. Bei Müdigkeit liegt die beste Diagnosegüte bei einem Trennwert von einem Skalenwert über der Baseline vor. In Hinblick auf die Aufmerksamkeit kann die Übereinstimmung zwischen subjektiver Aufmerksamkeit und der RASMUS-Diagnose durch eine Veränderung des Trennwerts nicht verbessert werden.

Tabelle 52. Ergebnisse der ROC-Kurvenanalyse bei Verwendung der Baseline und dem bestmöglichen Trennwert zur Diskriminierung zwischen kritischen und unkritischen Nutzerzuständen (Experiment 4)

	Beanspruchung		Müdigkeit		Aufmerksamkeit	
Trennwert subjektives Maß	BL	BL+3	BL	BL+1	BL	Level-1-SA
AUC	.630	.717**	.693*	.812*	.532	.72**
Sensitivität	48,3%	69%	41,7%	66,6%	63%	83,3%
Spezifität	77,8%	74,5%	96,9%	95,7%	30,6%	60,7%

* $p<.05$; ** $p<.01$

Insgesamt zeigt sich, dass die RASMUS-Diagnose nach Anpassung der Trennwerte hohe AUC-Werte aufweist, die signifikant von dem zu erwartenden Ergebnis bei zufälliger Klassifikation (AUC-Wert 0.5) abweichen und für eine hohe Genauigkeit der Klassifikation sprechen[8]. Bei der Aufmerksamkeit ist zu beachten, dass dies nur für die Verwendung von Level-1-SA als Vergleichsmaß gilt.

7.3 Bewertung der Ergebnisse

Im Folgenden werden die Ergebnisse aus dem Validierungsexperiment in Hinblick auf die drei diagnostizierten Nutzerzustände näher diskutiert. Außerdem werden mögliche Erklärungen für Leistungseinbrüche ohne Diagnoseergebnis erörtert und Einfluss- und Störfaktoren diskutiert.

7.3.1 Diagnose kritischer Beanspruchung

In Bezug auf Hypothese H1 zeigten die Ergebnisse der hypothesenprüfenden Untersuchung, dass die subjektive Anstrengung bei dem Diagnoseergebnis *kritische Beanspruchung* erwartungskonform höher bewertet wurde als bei der Baseline und auch signifikant höher bewertet wurde als bei Leistungseinbrüchen, bei denen die Beanspruchung von RASMUS als unkritisch klassifiziert wurde. Bei Betrachtung der Ergebnisse auf Individualebene (Anhang E.3, Abbildung 57) wird ersichtlich, dass die Anstrengung bei den meisten Personen sowohl bei Leistungseinbrüchen mit kritischer als auch mit unkritischer Beanspruchung höher bewertet wird als bei der Baseline und die Bewertungen teilweise nur geringfügig voneinander abweichen. Dies lässt sich damit erklären, dass sich die meisten Leistungseinbrüche in der Phase hoher Belastung ereignet haben, in der die Beanspruchung von den Teilnehmern vermutlich als vergleichsweise höher – relativ zum Baselinezustand – empfunden wurde. Wie aus Abbildung 47 hervorgeht, liegt der Zeitanteil, in der RASMUS die Beanspruchung in dieser Phase als kritisch bewertete, hingegen nur bei 37%.

Da viele Leistungseinbrüche, bei denen RASMUS keine kritische Beanspruchung diagnostiziert hat, mit einer positiven Abweichung der subjektiven Anstrengung von der Baseline verbunden sind, ergibt sich in der Post hoc-Analyse eine hohe Fallzahl falsch negativer Diagnoseentscheidungen (vgl. Tabelle 51). Diese geht mit einer vergleichsweise niedrigen Sensitivität und einem geringen negativen prädiktiven Wert einher. Allerdings zeigte sich in der ROC-Kurvenanalyse, dass die Übereinstimmung zwischen dem dichotomisierten subjektiven Urteil und der RASMUS-Diagnose deutlich erhöht werden kann, wenn ein höherer Trennwert (BL+3) für die Dichotomisierung gewählt wird. Dies kann so interpretiert werden, dass RASMUS die Beanspruchung mit höherer Wahrscheinlichkeit als kritisch klassifiziert, wenn diese in stärkerem Maße von dem Normalzustand abweicht. Der signifikante AUC-Wert sowie die hohen Werte zu Sensitivität (69%) und Spezifität (74,5%) bei Verwendung des Trennwerts BL+3 belegen, dass RASMUS in der Lage ist, Zustände kritischer und unkritischer Beanspruchung, wie sie durch die dichotomisierte subjektive Bewertung angezeigt werden, verlässlich zu erkennen.

[8] Nach Simundic (2008) kann die Genauigkeit bei einem AUC-Wert >.7 als gut bewertet werden.

7.3.2 Diagnose kritischer Müdigkeit

Leistungseinbrüche mit dem Diagnoseergebnis *kritische Müdigkeit* sind nur in wenigen Fällen und – wie die Individualanalyse zeigt – bei nur vier Personen aufgetreten. Dies schränkt die Aussagekraft der Ergebnisse ein. Dennoch konnten in der statistischen Analyse positive Abweichungen von der Baseline und signifikante Abweichungen zu Leistungseinbrüchen mit unkritischer Müdigkeit festgestellt werden. In der Post hoc-Untersuchung weist die RASMUS-Diagnose bereits bei Verwendung der Baseline als Trennwert einen AUC-Wert von knapp .7 und damit eine gute Übereinstimmung mit der subjektiven Bewertung auf. Bei Verwendung des Trennwerts „BL+1“ liegt mit einem AUC-Wert von über .8 eine sehr hohe Übereinstimmung vor.

Bei Betrachtung der Ergebnisse auf Individualebene zeigt sich allerdings, dass eine Versuchsperson (VP 3) vier Leistungseinbrüche mit kritischer Müdigkeit aufweist (vgl. Abbildung 56, Anhang E.3). Die Ergebnisse zur Diagnose kritischer Müdigkeit (richtig und falsch Positive) wurden daher maßgeblich durch die Bewertung dieser Person beeinflusst. Es ist daher fraglich, ob das Ergebnis auch auf andere Personen generalisierbar ist, für die kein Leistungseinbruch mit kritischer Müdigkeit diagnostiziert wurde.

Demgegenüber basieren die Ergebnisse zur Diagnose unkritischer Müdigkeit auf einer sehr hohen Fallzahl (n=69) und sind daher als aussagekräftiger einzuschätzen. Hier kann konstatiert werden, dass nahezu alle Leistungseinbrüche mit unkritischer Müdigkeit mit einer negativen Abweichung der subjektiven Müdigkeit von der Baseline korrespondieren (der negative prädiktive Wert liegt bei 90%). Außerdem wird bei negativen Abweichungen zu nahezu 100% ein unkritischer Zustand diagnostiziert (Spezifität: 97%). Dies ist auch bei Betrachtung der Ergebnisse auf Individualebene (vgl. Abbildung 57, Anhang E.3) ersichtlich.

Insgesamt weisen diese Ergebnisse darauf hin, dass RASMUS den Zustand der unkritischen Müdigkeit sehr verlässlich diagnostizieren kann. Die Ergebnisse deuten auch darauf hin, dass RASMUS in der Lage ist, Zustände kritischer Müdigkeit zu erkennen, wobei die Generalisierbarkeit dieser Aussage in Anbetracht der geringen Fallzahl fraglich ist. Gestützt wird diese Annahme allerdings zusätzlich durch die deskriptive Analyse der Auftretenshäufigkeit passiver aufgabenbezogener Müdigkeit pro Phase des Szenarios (vgl. Abbildung 47 in Abschnitt 7.2.1). So diagnostizierte RASMUS passive aufgabenbezogene Müdigkeit in der Monotoniephase, die diesen Zustand hervorrufen sollte, in mehr als der Hälfte der Zeit. Offensichtlich beeinträchtigte dieser Zustand die Leistungsfähigkeit der Versuchsteilnehmer jedoch zumeist nicht stark genug, um einen Leistungseinbruch auszulösen.

7.3.3 Diagnose kritischer Aufmerksamkeit

Die Ergebnisse zu dem Diagnoseergebnis *kritische Aufmerksamkeit* sind bezogen auf die Teilhypothesen H3a und H3b sehr unterschiedlich ausgefallen. Bei Verwendung der subjektiven Aufmerksamkeit als Vergleichsmaß können die Annahmen nicht bestätigt werden. Die Auswertung zum Level-1-SA spricht hingegen für die Validität der Echtzeitdiagnose. Hinsichtlich der subjektiven Aufmerksamkeit fällt bei Betrachtung der Ergebnisse auf Individualebene (vgl. Abbildung 57, Anhang E.3) auf, dass sich die Probanden interindividuell stark in der Bewertung unterscheiden, was sich in teilweise positiven und teilweise negativen Abweichungen von der

Baseline zeigt. Dagegen fallen die Unterschiede zwischen den Diagnoseergebnissen nur gering aus. Dies deutet darauf hin, dass die Probanden Probleme hatten, ihre Aufmerksamkeit einzuschätzen.

Zudem ist zu bedenken, dass die Einschätzung auch dadurch verzerrt sein kann, dass den Probanden gar nicht bewusst war, dass sie Aufgaben übersehen haben. Die Auswertung von Hypothese H3b bezüglich der Verwendung von Level-1-SA als Vergleichsmaß bestätigt, dass die Aufgabe, die den Leistungseinbruch auslöste, signifikant häufiger übersehen wurde, wenn von RASMUS eine kritische Aufmerksamkeit diagnostiziert wurde. Dies spiegelt sich auch in hohen diagnostischen Kennwerten in der Post hoc-Untersuchung wider. Im Vergleich zur subjektiven Bewertung stellt die Erfassung der ersten Ebene des Situationsbewusstseins ein objektives Maß dar, das, wie in Abschnitt 3.2 erläutert wurde, eng mit der Aufmerksamkeit verbunden ist. Es scheint daher als Vergleichsmaß besser geeignet zu sein als die subjektive Aufmerksamkeit.

7.3.4 Leistungseinbrüche ohne Diagnoseergebnis

In 11 Fällen traten Leistungseinbrüche auf, ohne dass RASMUS dafür einen kritischen Nutzerzustand identifizierte. Für diese Leistungseinbrüche wurde daher in einer Zusatzauswertung (vgl. Anhang E.3, Abbildung 58 und Abbildung 59) analysiert, inwiefern die subjektiven Bewertungen auf einen kritischen Nutzerzustand hinweisen, der durch RASMUS nicht erkannt wurde. Aus Abbildung 58 geht hervor, dass hinsichtlich der subjektiven Anstrengung in einem Fall (VP 7) eine Abweichung von der Baseline > 3 aufgetreten ist, die auch nach Anheben des Trennwerts auf Basis der ROC-Analyse eine kritisch hohe Beanspruchung anzeigt. Müdigkeit weicht bei drei Fällen, die sich auf die gleiche Person beziehen (VP 8), positiv von der Baseline ab und könnte somit bei dieser Person für den Leistungseinbruch verantwortlich gewesen sein. Die Aufmerksamkeit ist auf Basis des Level-1-SA in drei Fällen als kritisch zu bewerten.

Als weitere Ursachen für Leistungseinbrüche ohne Diagnoseergebnis wurden die nicht erfassten Nutzerzustände *Motivation* und *emotionaler Zustand* untersucht (vgl. Abbildung 59). Es zeigt sich, dass die Motivation bei diesen Leistungseinbrüchen sogar höher bewertet wurde als bei der Baseline und somit keinen Problemzustand darstellte. Aus den Bewertungen des SAM geht hingegen hervor, dass die Valenz teilweise geringer und die Erregung höher bewertet wurde als bei der Baseline, was auf ungünstige emotionale Zustände, wie Frustration, hinweist. Dafür sprechen auch positive Abweichungen in Hinblick auf die Dominanz (hohe Werte entsprechen bei dieser Skala einem Gefühl geringer Dominanz).

Aufgrund der Beobachtungen während des Experiments kann auch vermutet werden, dass Personen z.B. aufgrund geringer Erfahrung mehr Zeit als vorgesehen für die Aufgabenbearbeitung benötigten, ohne dass sich dies in einem kritischen Nutzerzustand äußerte. Unter diesem Aspekt erscheint es sinnvoll, individuelle Faktoren, auch wenn diese nur singulär erfasst werden, künftig in die Echtzeitdiagnose mit einzubeziehen.

Einige Leistungseinbrüche können aber vermutlich auch auf Ungenauigkeiten in der Bearbeitung zurückgeführt werden. Zum Beispiel beobachteten die Versuchsleiter, dass Teilnehmer die NRTT-Aufgabe zwar bearbeiteten, aber manchmal vergaßen, diese durch Klick auf den „Erledigen“-Button zu beenden, so dass diese Aufgabe von RASMUS als „nicht bearbeitet“ gewertet wurde und einen Leistungseinbruch auslösen konnte.

7.3.5 Einflussfaktoren auf die Ergebnisse des Validierungsexperiments

Insgesamt muss bei der Intepretation der Ergebnisse berücksichtigt werden, dass positive sowie negative Befunde zur Diagnosefähigkeit von RASMUS durch weitere Faktoren beeinflusst werden. Da die subjektive Bewertung des Nutzerzustands als Referenz genutzt wurde, stellt die Fähigkeit der Nutzer, ihren aktuellen Zustand korrekt einzuschätzen, einen wichtigen Faktor dar. In einigen Studien (z. B. Schmidt et al., 2009; Larue et al. 2011) wurde festgestellt, dass Nutzer ihre Müdigkeit und Aufmerksamkeit nicht immer korrekt einschätzen können. Eine schlechte Selbsteinschätzung der Probanden kann die Befunde somit maßgeblich beeinflusst haben.

Ein weiterer Einflussfaktor auf die Analyseergebnisse besteht darin, dass bei den einzelnen Versuchsteilnehmern unterschiedlich oft Leistungseinbußen aufgetreten sind. Somit sind ihre subjektiven Bewertungen in der Analyse unterschiedlich stark vertreten und etwaige Antworttendenzen der Teilnehmer können unterschiedlich stark ins Gewicht fallen. Die generelle Tendenz eher niedrig oder eher hoch zu bewerten, konnte ausgeglichen werden, indem nur die Abweichungen von der Baseline betrachtet wurden. Jedoch kann diese nicht ausgleichen, wenn Probanden bei Mehrfachbefragungen dazu tendieren, gleiche oder ähnliche Werte anzugeben, während andere dazu tendieren, die volle Skala auszunutzen. Die Auswertung auf Individualebene belegt die Relevanz interindividueller Unterschiede (vgl. Abbildung 57, Anhang E.3).

7.3.6 Resümee

Die Ergebnisse des Validierungsexperiments, die in Tabelle 53 zusammengefasst sind, weisen auf eine größtenteils hohe Übereinstimmung der Diagnoseergebnisse von RASMUS mit den jeweiligen Vergleichsmaßen hin. Dies spricht, insbesondere in Anbetracht dessen, dass auch die als Referenz genutzte subjektive Bewertung fehlerbehaftet ist und den wahren Zustand möglicherweise nicht exakt wiedergibt (vgl. Abschnitt 7.3.5), für eine valide Erfassung der drei untersuchten Nutzerzustände. Eine Einschränkung ergibt sich, da das Diagnoseergebnis für Aufmerksamkeit nicht in erwarteter Weise mit der subjektiven Bewertung korrespondiert. Wie in Abschnitt 7.3.3 erörtert wurde, ist hierbei jedoch eine eingeschränkte Validität des subjektiven Vergleichsmaßes wahrscheinlich, zumal die Diagnose kritischer Aufmerksamkeit durch RASMUS eine hohe Übereinstimmung mit dem alternativen Vergleichsmaß Level-1-SA aufweist.

Tabelle 53. Übersicht über die Ergebnisse des Validierungsexperiments (Experiment 4)

	Diagnose RASMUS	Vergleichsmaß	Ergebnis hypothesen-prüfende Untersuchung	Ergebnis zur Diagnosegüte
H1:	Hohe Beanspruchung	NASA-TLX Skala Anstrengung	✓	✓
H2:	Passive aufgabenbezogene Müdigkeit	Subjektive Müdigkeit	✓	✓
H3:	Falsche Aufmerksamkeits-verteilung	Subj. Aufmerksamkeit	X	X
		Level-1-SA	✓	✓

Legende: ✓ = erwartungskonform, X = nicht erwartungskonform

Kritisch anzumerken ist, dass die Ergebnisse nur auf einer kleinen Stichprobe von N=12 basieren und die Fallzahlen für Leistungseinbrüche mit kritischen Nutzerzuständen aufgrund des quasiexperimentellen Designs insbesondere mit passiver aufgabenbezogener Müdigkeit nur sehr gering ausfallen. Dies schließt die Berechnung komplexer Modelle und damit die Berücksichtigung individueller Unterschiede als Einflussfaktor in der Analyse aus. Kleine Stichproben schränken außerdem die Teststärke und auch die Übertragbarkeit der Ergebnisse ein (Tabachnick & Fidell, 2007).

Für den vorliegenden Anwendungskontext sind diese Limitierungen jedoch weniger maßgeblich. Einschränkungen der Teststärke erschweren lediglich die Ablehnung der Nullhypothese und damit die Bestätigung eines erwarteten Effekts. Bedeutsamer sind die individuellen Unterschiede und die Frage der Übertragbarkeit. Zur Übertragbarkeit ist anzumerken, dass das Validierungsexperiment mit dem Ziel durchgeführt wurde, zu prüfen, ob eine valide multifaktorielle Echtzeitdiagnose durch RASMUS in der derzeitig umgesetzten Form möglich ist. Der Aspekt der Übertragbarkeit der Ergebnisse auf andere Personen und Situationen stand dabei noch nicht im Fokus. Bezüglich individueller Unterschiede zeigte sich in den Analysen auf Individualebene, dass die Übereinstimmung mit der subjektiven Bewertung je nach Person unterschiedlich ausfallen kann. Es ist daher möglich, dass die Diagnoseergebnisse im Einzelfall nicht immer mit dem subjektiven Empfinden korrespondieren. Dazu ist anzumerken, dass die Diagnoseergebnisse zu kritisch ausgeprägten Nutzerzuständen in dem zu entwickelnden adaptiven System lediglich für die Auswahl geeigneter Adaptierungsstrategien herangezogen werden (vgl. Abschnitt 6.1). Eine möglicherweise nicht adäquate Diagnose in einzelnen Fällen ist somit mit keinen gravierenden Konsequenzen verbunden. Insgesamt sprechen die Ergebnisse daher dafür, dass die Echtzeitdiagnose in der derzeitigen Umsetzung eingesetzt werden kann, um in einem adaptiven System Zustände *hoher Beanspruchung*, *passiver aufgabenbezogener Müdigkeit* und einer *falschen Aufmerksamkeitsverteilung* zu erfassen.

7.4 Weiterführende Arbeiten zur Optimierung der Diagnose

Die Ergebnisse aus dem Validierungsexperiment weisen darauf hin, dass die Diagnoseergebnisse von RASMUS bereits überwiegend mit den Phasen im Szenario (vgl. Abschnitt 7.2.1) und den subjektiven Bewertungen korrespondieren (vgl. Abschnitte 7.2.2 und 7.2.3). Die Formulierung der im Experiment verwendeten Regeln erfolgte allerdings vorwiegend auf Basis der Erkenntnisse aus Experiment 3 (vgl. Abschnitt 6.5), das noch keine Echtzeitdiagnose beinhaltete, und in der die Indikatoren *Atemfrequenz* und *HRV* noch nicht erfasst worden waren. Es ist daher wahrscheinlich, dass die Diagnoseregeln anhand der Daten aus dem Validierungsexperiment noch weiter optimiert werden können.

In Hinblick auf die kritischen Nutzerzustände *hohe Beanspruchung* und *passive aufgabenbezogene Müdigkeit* deuten die in Anhang E.3, Abbildung 61 dargestellten Ergebnisse darauf hin, dass einige Indikatoren (z.B. HRV) zum Zeitpunkt von Leistungseinbrüchen sehr viel häufiger kritisch ausgeprägt sind als andere (z.B. Pupillenweite). In einer weiterführenden Arbeit, die nicht Bestandteil der vorliegenden Dissertation ist, wurde daher unter Verwendung der Daten aus dem Validierungsexperiment beispielhaft für den Nutzerzustand Beanspruchung untersucht, welche

Grenzwerte für kritische Ausprägungen der physiologischen Indikatoren besser geeignet sein könnten, um zwischen kritischen und unkritischen Zuständen (basierend auf dem subjektiven Urteil) zu diskriminieren (Bruder & Schwarz, 2019). Als Verfahren wurde die in Abschnitt 7.2.3 vorgestellte ROC-Kurvenanalyse verwendet. Die auf diese Weise ermittelten modifizierten Diagnoseregeln wurden in einer Replikation des Validierungsexperiments mit *N*=15 Probanden auf ihre praktische Anwendbarkeit geprüft. Hierbei zeigte sich, dass die ursprünglichen Regeln entgegen den Erwartungen eine bessere Diagnosegüte aufweisen als die modifizierten Regeln. Die Ergebnisse aus dem Validierungsexperiment, die in den Abschnitten 7.2.2 und 7.2.3 dargestellt wurden, konnten bei Verwendung der ursprünglichen Regeln im Wesentlichen repliziert werden. Es erscheint folglich ratsam, die ursprünglichen Grenzwerte beizubehalten.

Es bieten sich jedoch noch weitere Möglichkeiten zur Optimierung der Diagnose: Da das Intervall, über das der gleitende Mittelwert berechnet wird, nach eigenem Ermessen auf 30 Sekunden festgelegt wurde, stellt sich die Frage, inwiefern kürzere oder längere Intervalle besser mit dem zum Zeitpunkt des Leistungseinbruchs vorliegenden Nutzerzustand korrespondieren. Außerdem ist zu prüfen, ob ein längeres Intervall zur Ermittlung der Baseline (derzeit 60 Sekunden) genutzt werden sollte, um den Baselinezustand stabiler zu erfassen.

Darüber hinaus könnte die Robustheit der Diagnose durch Hinzunahme weiterer Sensoren gesteigert werden. Derzeit werden die physiologischen Parameter *HRV* und *Atemfrequenz* durch den BioHarness 3 erfasst. Kommt es während der Messwerterfassung bei diesem Sensor zu einer Störung oder einem Ausfall, stehen beide Maße nicht mehr für die Nutzerzustandsdiagnose zur Verfügung. Zudem zeigte sich in den Analysen zum Stand der Forschung (vgl. Abschnitt 2.3.3), dass die HRV und die Atemfrequenz nicht voneinander unabhängig sind, so dass auf beide Maße ähnliche Störvariablen (z.B. Sprechen, körperliche Aktivität) wirken. In weiterführenden Arbeiten soll daher die Eignung weiterer Sensoren und Parameter zur Steigerung der Robustheit der Diagnose empirisch untersucht werden.

Die Diagnose einer *falschen Aufmerksamkeitsverteilung* basiert derzeit auf Informationen darüber, welche Aufgaben zu bearbeiten sind und mit welcher Aufgabe sich der Nutzer beschäftigt. Eine Aufgabe wird dann als in Bearbeitung befindlich erkannt, wenn das betreffende Objekt mit der Maus ausgewählt wurde. Allerdings zeigten die Beobachtungen im Validierungsexperiment, dass das zuletzt ausgewählte Objekt nicht zwingend das Objekt ist, auf das der Teilnehmer seine Aufmerksamkeit gerichtet hat. Um die Genauigkeit der Diagnose zu erhöhen, könnte es daher sinnvoll sein, die über den Eyetracker erfassten Fixationen einzubeziehen und zu prüfen, welches Objekt der Teilnehmer gerade fixiert (vgl. Vorgehen von Bosse et al., 2009, beschrieben in Abschnitt 2.2.5). Dabei ist jedoch zu bedenken, dass es sich bei den Luftkontakten um bewegliche Objekte geringer Größe handelt (so genannte „dynamische Areas of Interest - AOI“; vgl. Papenmeier & Huff, 2010), die sich überlagern können. Inwiefern eine zuverlässige Bestimmung der fixierten dynamischen AOI bei der verwendeten Experimentalaufgabe möglich ist, ist in künftigen Analysen zu klären.

Ein anderer Aspekt, der im Validierungsexperiment weniger im Fokus stand, betrifft die Diagnose von Leistungseinbrüchen. Die Regeln für Leistungseinbrüche basieren auf der Zeitdauer, die für die Bearbeitung der Teilaufgaben zur Verfügung steht. Für die Festlegung der Zeitdauern wurden

zunächst die Dringlichkeit der Aufgabe und die aus Experiment 3 abgeleitete durchschnittliche Bearbeitungszeit zugrunde gelegt. Es stellt sich nun die Frage, ob diese Regeln auf Basis der Erkenntnisse aus dem Validierungsexperiment modifiziert werden sollten. Die Analyse zu den Häufigkeiten der Leistungseinbrüche pro Person (vgl. Abschnitt 7.2.1) zeigte, dass ein Teilnehmer in der Lage war, das Szenario mit nur einem Leistungseinbruch zu bearbeiten. Dies weist darauf hin, dass die zur Verfügung stehende Bearbeitungsdauer nicht zu knapp gewählt wurde. Im Rahmen der weiteren Forschungsarbeiten zum Adaptierungsmanagement ist zu prüfen, ob eine Verkürzung der Bearbeitungsdauern für eine frühzeitigere Unterstützung des Nutzers angebracht ist.

Außerdem wurde beobachtet, dass Teilnehmer die NRTT-Aufgabe in einigen Fällen erst spät gesehen haben. Da die Bearbeitung dieser Aufgabe einige Zeit in Anspruch nimmt, löste die Aufgabe oft während der Bearbeitung einen Leistungseinbruch aus. Es könnte daher zweckmäßiger sein, für das Auslösen von Adaptierungen statt der Zeit bis zur vollständigen Bearbeitung einer Aufgabe, die Zeit bis zum Bearbeitungsbeginn zugrunde zu legen.

8 Abschließende Diskussion der Forschungsarbeit und Ausblick

In der vorliegenden Forschungsarbeit wurde mit RASMUS (Real time Assessment of Multidimensional User State) ein generisches Konzept für eine multifaktorielle Echtzeitdiagnose des Nutzerzustands in adaptiver Mensch-Maschine-Interaktion entwickelt. Das Konzept wurde exemplarisch für einen Anwendungsfall aus dem Bereich der Luftraumüberwachung umgesetzt und empirisch validiert. In diesem Kapitel erfolgt eine abschließende Bewertung des Diagnosekonzepts (Abschnitt 8.1). Außerdem wird ein Ausblick auf potenzielle Erweiterungen und Einsatzmöglichkeiten der Echtzeitdiagnose gegeben (Abschnitt 8.2).

8.1 Bewertung des Diagnosekonzepts

Die Bewertung des Diagnosekonzepts ist in drei Abschnitte unterteilt. Abschnitt 8.1.1 bezieht sich auf die Frage, inwiefern die Konzeption und Umsetzung von RASMUS die Zielsetzungen des Promotionsvorhabens sowie die spezifischen Anforderungen, die sich aus der Literaturanalyse ergeben haben, erfüllt. In Abschnitt 8.1.2 werden die Erkenntnisse aus dem Validierungsexperiment herangezogen, um zentrale Elemente des Diagnosekonzepts zu bewerten. Zuletzt vergleicht Abschnitt 8.1.3 das Diagnosekonzept mit einem ähnlichen multifaktoriellen Ansatz aus dem Luftfahrtbereich.

8.1.1 Bezug der konzeptionellen Maßnahmen zu Zielsetzungen und Anforderungen

Das Diagnosekonzept von RASMUS sieht eine Erfassung des Nutzerzustands *in Echtzeit* und auf *individueller Ebene* vor. Neuartig an dem Konzept ist, dass der Nutzerzustand *multidimensional* als Zusammenwirken von sechs Zustandsdimensionen betrachtet sowie *multifaktoriell* – unter Berücksichtigung der Einflussfaktoren auf den Nutzerzustand – erfasst und bewertet wird. Das Diagnosekonzept erfüllt damit die Anforderungen, die sich aus der übergeordneten Zielsetzung einer ganzheitlichen Betrachtung und Adressierung von Problemzuständen in adaptiver Mensch-Maschine-Interaktion ergeben haben (vgl. Abschnitt 1.5).

Neben diesen grundlegenden Anforderungen ergaben sich aus den Analysen zum Stand der Forschung (Kapitel 2) weitere Anforderungen, die bei der Konzeption und der exemplarischen Umsetzung von RASMUS zu berücksichtigen waren (vgl. Abschnitt 2.5, Tabelle 7). Tabelle 54 stellt dar, welche Merkmale von RASMUS auf welche Anforderungen Bezug nehmen. Insgesamt zeigt sich, dass alle acht Anforderungen bei der Konzeption und Umsetzung von RASMUS adressiert werden konnten.

Darüber hinaus wurden bei der Analyse psychologischer Theorien und Modelle in Kapitel 3 Erkenntnisse gewonnen, die für die Diagnose und Adaptierung von Nutzerzuständen relevant sind. Das daraus abgeleitete generische Modell zum Nutzerzustand (vgl. Abbildung 15, S. 65) bildet die Grundlage für eine multifaktorielle Nutzerzustandsbewertung. Weitere Aspekte, wie die Differenzierung zwischen verschiedenen Ausprägungsformen einer Zustandsdimension sollten in eine künftige Weiterentwicklung von RASMUS mit einfließen, um die Nutzerzustandsdiagnose zu erweitern und zu präzisieren (vgl. Abschnitt 8.2.1).

Tabelle 54. Berücksichtigung der Anforderungen aus der Literaturanalyse (Kapitel 2) im Diagnosekonzept von RASMUS

1	**Störeinflüsse**	• Das Diagnosekonzept sieht vor, mögliche Störeinflüsse als Einflussfaktoren in die Diagnose einzubeziehen.
2	**Reliabilität und zeitliche Stabilität**	• Mangelnde Reliabilität einzelner Indikatoren wird durch die Kombination mehrerer Indikatoren ausgeglichen (ein Nutzerzustand wird nur dann als kritisch diagnostiziert, wenn die Mehrzahl der Indikatoren eine kritische Ausprägung anzeigt). • Aufgrund mangelnder zeitlicher Stabilität der Indikatoren auf Individualebene wird auf eine nutzerspezifische Auswahl von Indikatoren verzichtet.
3	**Interindividuelle Unterschiede**	• Zur Bewertung kritischer Ausprägungen der physiologischen Maße werden die Abweichungen von einer individuellen Baseline herangezogen.
4	**Menschliche Selbstregulierung**	• Unterstützungsbedarf wird erst bei Auftreten von Leistungseinbrüchen ermittelt, um produktive Selbstregulierungsstrategien des Menschen nicht zu unterbinden.
5 **6**	**Ursachenanalyse** **Kontextfaktoren**	• In die Diagnose werden Einflussfaktoren auf den Nutzerzustand einbezogen, die Kontextinformationen liefern und als Ursachen für kritische Nutzerzustände in Betracht kommen.
7	**Oszillieren der Adaptierung**	• Ein Oszillieren physiologischer Maße wird durch die Berechnung gleitender Mittelwerte reduziert. • Ein Oszillieren der Adaptierung wird durch die Verwendung von Leistungseinbrüchen als Auslöser für Adaptierungen und das Einführen einer „Totzeit“ unterbunden.
8	**Künstlichkeit von Laboraufgaben**	• Die Echtzeitdiagnose wurde für eine maritime Luftraumüberwachungsaufgabe umgesetzt, die engen Bezug zu realen Aufgaben im Bereich AAW hat.

Anmerkung: Die erste Spalte weist auf die in Tabelle 7, Abschnitt 2.4 aufgeführten Anforderungen hin.

8.1.2 Erkenntnisse aus dem Validierungsexperiment

Die Ergebnisse aus dem Validierungsexperiment wurden bereits in Abschnitt 7.3 erörtert. Sie weisen darauf hin, dass RASMUS in der derzeitigen Umsetzung in der Lage ist, den Nutzerzustand multidimensional in Hinblick auf die drei potenziell kritischen Zustände *hohe Beanspruchung*, *passive aufgabenbezogene Müdigkeit* und *falsche Aufmerksamkeitsverteilung* zu bewerten. In Anbetracht der geringen Fallzahlen insbesondere für *passive aufgabenbezogene Müdigkeit* erscheint es allerdings ratsam, die Diagnosefähigkeit in weiteren Untersuchungen zu prüfen. Überdies wurden in Abschnitt 7.4 Möglichkeiten zur Verbesserung der Diagnosefähigkeiten identifiziert, die in künftigen Untersuchungen ebenfalls zu evaluieren sind.

Darüber hinaus gibt das Validierungsexperiment auch Aufschluss über die praktische Relevanz grundlegender Merkmale des Diagnosekonzepts. So weist die Tatsache, dass es bei den verschiedenen Versuchsteilnehmern unterschiedlich häufig zu Leistungseinbrüchen kam, darauf hin, dass es angebracht ist, die Bewertung von Unterstützungsbedarf individuell vorzunehmen und nur dann Unterstützung anzubieten, wenn der Nutzer diese benötigt.

Im Experiment zeigte sich auch, dass unterschiedliche Arten kritischer Nutzerzustände mit Leistungseinbrüchen assoziiert sein können und diese teilweise auch zeitgleich vorliegen können. Dies bekräftigt die Notwendigkeit einer multidimensionalen Bewertung des Nutzerzustands.

Außerdem konnte beobachtet werden, dass RASMUS zu einem hohen Prozentanteil potenziell kritische Nutzerzustände während der Laufzeit des Szenarios diagnostizierte (vgl. Abbildung 47). Eine Adaptierung auf Basis der Nutzerzustandsdiagnosen würde somit zu häufigen Adaptierungen führen, die produktiven Selbstregulierungsstrategien entgegenwirken könnten. Dies spricht für das Vorgehen, Adaptierungsbedarf auf Basis von Leistungsmaßen zu bestimmen.
Eine Zusatzauswertung, deren Ergebnisse in Anhang E.3, Abbildung 60 dargestellt sind, ergab zudem, dass in den meisten Fällen nur drei der fünf Indikatoren für hohe Beanspruchung und passive aufgabenbezogene Müdigkeit bei einem Leistungseinbruch kritisch ausgeprägt waren. Dies bestätigt den Ansatz, den Nutzerzustand durch eine Kombination verschiedener Indikatoren multifaktoriell zu bewerten, um Messprobleme bei einzelnen Indikatoren auszugleichen.
Die Erkenntnisse aus dem Validierungsexperiment weisen somit darauf hin, dass RASMUS adäquate Möglichkeiten bietet, um den identifizierten Herausforderungen bei der Nutzerzustands-diagnose in adaptiven Mensch-Maschine-Systemen zu begegnen.

8.1.3 Vergleich mit dem Forschungsvorhaben „Human Performance Envelope“

Im Rahmen des Forschungsprogramms *Horizon 2020* wird derzeit in einem internationalen Verbundprojekt der *Human Performance Envelope (HPE),* ein ebenfalls multifaktorieller Ansatz zur Bewertung der Leistungsfähigkeit von Operateuren im Bereich der Flugzeugführung, entwickelt (Silvagni et al., 2015). Da von diesem Ansatz erst nach Fertigstellung des Diagnosekonzepts von RASMUS Kenntnis gewonnen wurde, konnten die Arbeiten zum HPE nicht in die Konzepterstellung einbezogen werden. Durch einen Post hoc-Vergleich beider Ansätze sollen nun Unterschiede und Gemeinsamkeiten identifiziert werden, die zu einer Bewertung und möglichen Weiterentwicklung von RASMUS beitragen können.

Der Ansatz des HPE zielt darauf ab, Faktoren und Faktorenkombinationen zu identifizieren und zu erfassen, die sich auf die Leistung von Piloten im Cockpit auswirken, und daraus Erkenntnisse für die Verbesserung der Mensch-Maschine-Schnittstelle abzuleiten. Der HPE basiert auf den Arbeiten von Edwards (2013), die in ihrer Dissertation eine systematische Untersuchung von Einfluss-faktoren auf die Leistung von Fluglotsen vorgenommen hat und insbesondere auch Wechselwir-kungen zwischen diesen Faktoren untersuchte; einige Erkenntnisse daraus sind in Abschnitt 3.2 aufgeführt. Dabei stellten sich acht Faktoren als besonders relevant heraus: *Beanspruchung, Müdigkeit, Stress, Situationsbewusstsein, Aufmerksamkeit/Vigilanz, Team, Kommunikation* und *Vertrauen.*

Diese Faktoren finden weitgehend auch in RASMUS Berücksichtigung: Die ersten fünf Faktoren werden bei RASMUS im Rahmen des multidimensionalen Nutzerzustands betrachtet. *Stress* wird dabei nicht als eigene Zustandsdimension aufgeführt; Stresszustände werden allerdings als eine Ausprägung des emotionalen Zustands gesehen, die in Verbindung mit hoher Beanspruchung stehen (vgl. Abschnitt 3.1.2). Die Faktoren *Team* und *Kommunikation* können in dem generischen Modell zum Nutzerzustand (vgl. Abbildung 15, Abschnitt 3.4), das RASMUS zugrunde liegt, dem sozialen Kontext zugeordnet werden. Der soziale Kontext wurde jedoch bei der exemplarischen Umsetzung von RASMUS nicht näher betrachtet, da er für die verwendete Einzelaufgabe keine Relevanz hatte.

Vertrauen ist im Modell zum Nutzerzustand nicht explizit aufgeführt. Sofern sich das Vertrauen auf die Technik bezieht, wird es in starkem Maße durch die Zuverlässigkeit des technischen Systems beeinflusst (vgl. Abschnitt 1.2.2), die im Modell eine Eigenschaft des technischen Systems darstellt.

Ein Nutzerzustand, der hingegen im HPE nicht berücksichtigt wird, ist die *Motivation*. Vermutlich ist dies darauf zurückzuführen, dass der Problemzustand einer (zu) geringen Motivation im operativen Kontext nur selten als Problem wahrgenommen wird. Dies bestätigen auch eigene Befragungen mit Anwendern. Von hoher Relevanz ist die Motivation jedoch, wie sich in Abschnitt 2.2.3 zeigte, in Ausbildung und Training. Dies ist ein Bereich, in dem RASMUS ebenfalls Anwendung finden könnte (vgl. Abschnitt 8.2.2), so dass die Berücksichtigung der Motivation als Dimension des Nutzerzustands im eigenen Ansatz gerechtfertigt erscheint.

Unterschiede ergeben sich insbesondere aus dem anders gearteten Fokus beider Forschungsvorhaben. Während es im HPE vorwiegend um die Frage geht, wie die Leistung von Piloten durch die Gestaltung der Mensch-Maschine-Schnittstelle im Cockpit verbessert werden kann, zielt der eigene Ansatz darauf ab, die Effektivität und Sicherheit von Mensch-Maschine-Systemen durch ein adaptives Verhalten des technischen Systems zu verbessern. Dies impliziert, dass eine Echtzeitdiagnose relevanter Nutzerzustände und Einflussfaktoren, wie sie im Rahmen der vorliegenden Dissertation entwickelt und validiert wurde, notwendig ist. Im HPE erfolgt die Analyse der betrachteten Faktoren vorwiegend „offline“ nach der Datenerhebung. Eine Echtzeiterfassung wurde – nach eigenem Kenntnisstand – bislang noch nicht umgesetzt.

Außerdem stellt der eigene Ansatz ein generisches Diagnosekonzept dar, das nicht auf eine konkrete Anwendungsdomäne zugeschnitten ist. Dieses basiert auf einer detaillierten Betrachtung von psychologischen Theorien und Modellen, die eine strukturierte Darstellung relevanter Faktoren in einem generischen Modell zum Nutzerzustand ermöglichte. Der HPE kann demgegenüber durch die Fokussierung auf eine konkrete Anwendungsdomäne tiefergehend zu berücksichtigende Faktoren und Wechselwirkungen für diesen Bereich untersuchen.

Insgesamt weisen die Übereinstimmungen dieser unabhängig voneinander entwickelten und durchgeführten Forschungsvorhaben darauf hin, dass der Ansatz einer mehrdimensionalen Betrachtung des Nutzerzustands und einer umfassenden Berücksichtigung von Einflussfaktoren auf die Leistung ein probater Weg sein kann, um mit den Herausforderungen der Mensch-Maschine-Interaktion in Zukunft umzugehen. Inhaltlich bestätigen die Erkenntnisse zum HPE größtenteils die praktische Relevanz der im Promotionsvorhaben betrachteten Zustandsdimensionen für den Luftfahrtbereich. Bei einer künftigen Weiterentwicklung und möglichen Anwendung von RASMUS auf die Luftfahrt empfiehlt es sich, die Erkenntnisse des HPE einzubeziehen.

8.2 Ausblick

Zuletzt sei auf Möglichkeiten hingewiesen, wie die Echtzeitdiagnose RASMUS in der derzeitig umgesetzten Form erweitert werden könnte, und für welche weiteren Anwendungsbereiche und Einsatzzwecke das Diagnosekonzept genutzt werden könnte.

8.2.1 Erweiterung der Diagnosefähigkeiten

Zusätzlich zu den drei derzeitig erfassten und experimentell untersuchten kritischen Nutzerzuständen *hohe Beanspruchung*, *passive aufgabenbezogene Müdigkeit* und *falsche Aufmerksamkeitsverteilung* ist eine Diagnose weiterer Zustandsdimensionen durch RASMUS denkbar. In Hinblick auf den *emotionalen Zustand* könnten die von der iMotions-Komponente „FACET“ durchgeführten Klassifizierungen zum emotionalen Zustand in die Diagnose einbezogen werden. Diese Software analysiert den über eine Webcam erfassten Gesichtsausdruck und bewertet in Echtzeit den Ausprägungsgrad von neun verschiedenen Gefühlszuständen sowie der positiven und negativen Valenz. Da diese Daten im Validierungsexperiment bereits aufgezeichnet wurden, besteht die Möglichkeit post hoc zu untersuchen, inwiefern die Klassifizierungen zum Zeitpunkt von Leistungseinbrüchen mit den subjektiven Bewertungen des emotionalen Zustands übereinstimmen.

Bei der *Motivation* zeigte sich in Abschnitt 2.2.3, dass die Nutzerinteraktion Aufschluss über den motivationalen Zustand geben kann. Auf dieser Grundlage kann vermutet werden, dass in dem verwendeten Experimentalparadigma ein Zustand geringer Motivation vorliegt, wenn die *Anzahl Aufgaben* kritisch hoch ist, während die *Anzahl Mausklicks* kritisch niedrig ist (dies weist darauf hin, dass viele Aufgaben zu bearbeiten sind, der Nutzer aber wenig Aktivität zeigt, um die Zahl der Aufgaben zu reduzieren). Um diese Annahme zu überprüfen, wurde die Motivation der Teilnehmer bereits im Validierungsexperiment auf Basis einer entsprechenden Diagnoseregel bewertet. Allerdings konnte bei keinem der Teilnehmer eine kritisch niedrige Motivation bei Leistungseinbrüchen festgestellt werden. Dieser Befund stimmt mit den subjektiven Bewertungen überein (die Motivation wurde stets als „hoch“ oder „sehr hoch“ bewertet). Die Frage, ob mit dieser Regel Zustände geringer Motivation erkannt werden können, bleibt jedoch offen.

In Bezug auf das *Situationsbewusstsein* konnte festgestellt werden, dass die Diagnose einer falschen Aufmerksamkeitsverteilung eng mit dem Level-1-SA (Endsley,1995) verbunden ist (vgl. Abschnitt 7.2.2). Eine Präzisierung könnte durch die in Abschnitt 2.2.6 beschriebene Kombination aus Fixationen und ERPs möglich sein. Für eine valide, kontinuierliche Erfassung des Situationsbewusstseins (insbesondere der Level 1 und 2) scheint allerdings noch weitere Forschung nötig.

Neben der Integration weiterer Zustandsdimensionen in die Diagnose könnte die Diagnose auch für die betrachteten drei Nutzerzustände erweitert werden. Zum Beispiel geht aus den Forschungsarbeiten von Wickens (1984) hervor, dass eine Unterscheidung zwischen perzeptiver, kognitiver und psychomotorischer Beanspruchung einen Mehrwert für zielgerichtete Adaptierungen bieten könnte (vgl. Abschnitt 3.1.1). Außerdem erscheint es angebracht, individuelle Faktoren, wie die Erfahrung oder das Wohlbefinden, in die Diagnose einzubeziehen. Diese Faktoren sind im Diagnosekonzept bereits als Indikatoren vorgesehen (vgl. Abschnitt 6.1). Bei der exemplarischen Umsetzung der Echtzeitdiagnose wurden sie zunächst noch nicht berücksichtigt, da sie bei einmaliger Erfassung auf individueller Ebene eine Konstante darstellen und keine Veränderungen des Nutzerzustands anzeigen können. Insbesondere variable individuelle Faktoren wie das Wohlbefinden könnten allerdings relevant für den wiederholten Einsatz (Sitzungen an unterschiedlichen Tagen) des Systems werden. Außerdem könnten sie bei Leistungseinbrüchen, bei denen die betrachteten Nutzerzustände keine kritische Ausprägung aufweisen, Aufschluss über zugrundeliegende Ursachen geben und die Auswahl geeigneter Adaptierungsstrategien unterstützen.

8.2.2 Einsatzmöglichkeiten der multifaktoriellen Echtzeitdiagnose

Zu Beginn dieser Dissertation wurde darauf hingewiesen, dass Unfälle von Mensch-Maschine-Systemen, wie das Air-France-Unglück, zu einem großen Anteil durch menschliche Fehler und kritische Nutzerzustände bedingt sind. Mit der multifaktoriellen Echtzeitdiagnose RASMUS wurde die Grundlage für eine dynamische Adaptierung des technischen Systems geschaffen, die eine bedarfsgerechte Unterstützung des Menschen – und damit eine Reduzierung kritischer Nutzerzustände und menschlicher Fehler ermöglichen kann. In dem übergeordneten adaptiven Gestaltungsansatz (vgl. Abschnitt 1.4) stellt RASMUS die Diagnosekomponente dar, die Informationen zu Leistungseinbrüchen, kritischen Nutzerzuständen sowie deren möglichen Ursachen in Echtzeit bereitstellt. Das Adaptierungsmanagement ADAM (vgl. Fuchs & Schwarz, 2017) verarbeitet diese Informationen weiter, um Unterstützungsbedarf zu identifizieren und Adaptierungsstrategien auszuwählen und zu konfigurieren, die geeignet sind, den diagnostizierten kritischen Zuständen effektiv entgegenzuwirken. ADAM wurde bereits konzeptionell entwickelt. Um die Effektivität des adaptiven Gesamtsystems untersuchen und demonstrieren zu können, soll ADAM ebenfalls exemplarisch für die betrachtete Luftraumüberwachungsaufgabe umgesetzt und in das bestehende Experimentalsystem integriert werden.

Da es sich bei der multifaktoriellen Echtzeitdiagnose RASMUS und dem Adaptierungsmanagement ADAM um generische Konzepte handelt, ist eine Übertragung auf jegliche Anwendungsdomänen denkbar, in denen kritische Nutzerzustände und menschliches Versagen die Effektivität und Sicherheit von Mensch-Maschine-Systemen gefährden können. Neben der Luftraumüberwachung zählen dazu beispielsweise auch die Fahrzeugführung, die Luftfahrt oder Leitzentralen. Zu beachten ist allerdings, dass Indikatoren und Regeln für die Diagnose und die Adaptierung, wie in Kapitel 6 beschrieben, auf den jeweiligen Anwendungsfall abgestimmt werden müssen.

Einsatzmöglichkeiten finden sich nicht nur im operativen Betrieb sicherheitskritischer Systeme sondern auch in den Bereichen Training und Ausbildung. Zum Beispiel wäre denkbar, dass RASMUS mentale Zustände, die für den Lernerfolg bedeutsam sind, erfasst. Dies würde es ermöglichen, das Training, z.B. in Simulatoren, adaptiv an den Zustand des Lernenden anzupassen und gewünschte Zustände gezielt hervorzurufen und zu überwachen. Der letztgenannte Aspekt ist insbesondere für das Training von Resilienz und Bewältigungsstrategien relevant, in denen kritische Nutzerzustände, wie Stress oder passive aufgabenbezogene Müdigkeit erzeugt werden müssen, um den erfolgreichen Umgang mit diesen Zuständen beüben zu können (vgl. Jones, Hale, Dechmerowski & Fouad, 2012).

Informationen zu Veränderungen des mentalen Zustands während des Trainings könnten auch herangezogen werden, um den Trainingserfolg zu bewerten. Zum Beispiel könnte mit Hilfe der Diagnoseergebnisse von RASMUS untersucht werden, ob das Training in der Lage ist, die Leistung zu verbessern und kritische mentale Zustände zu reduzieren. Schließlich ist auch eine Anwendung bei der Leistungsbewertung im Rahmen von Eignungstests oder Zertifizierungen denkbar. Insbesondere dort, wo die Bewertung weitgehend auf den subjektiven Urteilen der Ausbilder beruht (z.B. bei der Bewertung des Situationsbewusstseins von Prüflingen in einem Simulatortest), könnte die Justiziabilität durch die Berücksichtigung objektiver Diagnosebefunde von RASMUS erhöht werden.

9 Literatur

Adair, J. G. (1970). Pre-experiment attitudes toward psychology a determinant of subject behaviour. Paper presented at a symposium on *Methodological Problems in Research with Human Subjects* of the Canadian Psychological Asscociation, Winnipeg, Manitoba, May 1970.

Åkerstedt, T. (1990). Psychological and psychophysiological effects of shift work. *Scandinavian Journal of Work, Environment & Health, 16* (Suppl. 1), 67–73.

Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience, 52*(1-2), 29–37.

Åkerstedt, T., Mollard, R., Samel, A., Simons, M. & Spencer, M. (2003). *The role of EU FTL legislation in reducing cumulative fatigue in civil aviation.* Beitrag zum ETSC Meeting am 19. Februar 2003. Online verfügbar unter: https://www.eurocockpit.be/sites/default/files/Akerstedt-Mollard-Samel-Simons-Spencer-2003.pdf (letzter Zugriff: 22.12.2016)

Arciszewski, H. F. R., Greef, T. E. De, & Van Delft, J. H. (2009). Adaptive Automation in a Naval Combat Management System. *IEEE Transactions on Systems, Man, and Cybernetics 39*(6), 1188–1199.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130*(2), 224-237.

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological bulletin, 120*(3), 338-375.

Baccino, T., & Manunta, Y. (2005). Eye-fixation-related potentials: Insight into parafoveal processing. *Journal of Psychophysiology, 19*(3), 204–215.

Backs, R. W., & Ryan, A. M. (1992). Multimodal measures of mental workload during dual-task performance: Energetic demands of cognitive processes. *Proceedings of the Human Factors Society 36th Annual Meeting, Human Factors Society*, Santa Monica, CA, 1413-1417.

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task. *International Journal of Psychophysiology, 16*(1), 57–68.

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

Balaban, C. D., Prinkey, J., Frank, G., & Redfern, M. (2005). Postural measurements seated subjects as gauges of cognitive engagement. In D. D. Schmorrow (Ed.), *Proceedings of the 11th International Conference on Human-Computer Interaction (Augmented Cognition International)* (pp. 321–328). New York: Lawrence Erlbaum.

Barcelo, F., Gale, A., & Hall, M. (1995). Multichannel EEG power reflects information processing and attentional demands during visual orienting. *Journal of Psychophysiology, 9*, 32-44.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology, 42*(3), 361-377.

Barker, R. A., Edwards, R. E., O'Neill, K. R., & Tollar, J. R. (2004). *DARPA Improving Warfighter Information Intake Under Stress - Augmented Cognition Concept Validation Experiment (CVE).* Analysis Report for the Boeing Team. Arlington: DARPA.

Barker, R. A., & Edwards, R. E. (2005). The Boeing Team Fundamentals of Augmented Cognition. *Proceedings of the 1st International Conference on Augmented Cognition*, Las Vegas, NV, 22-27 July 2005.

Barr, L., Popkin, S., & Howarth, H. (2009). *An Evaluation of Emerging Driver Fatigue Detection Measures and Technologies*. Final Report FMCSA-RRR09-005. Federal Motor Carrier Safety Administration, Washington DC, USA.

BEA (2012). *Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF447 Rio de Janeiro – Paris*. Paris: Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. Online verfügbar unter: http://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf (letzter Zugriff: 17.12.2015).

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 142-162). New York: Cambridge University Press.

Beaumont, M., Burov, A., Carter, R., Cheuvront, S. N., Sawka, M. N., Wilson, G., et al. (2004). Risk Factors – Individual State (Chapter 3.2) In: NATO Research & Technology Organisation (ed.): *Operator Functional State Assessment. NATO-RTO-TR-HFM-104*. Neully-sur-Seine: NATO-RTO.

Behneman, A., Kintz, N., Johnson, R., Berka, C., Hale, K., Fuchs, S., et al. (2009). Enhancing Text-Based Analysis Using Neurophysiological Measures. In D. D. Schmorrow, I. V. Estabrooke & M. Grootjen (Eds.), *Foundations of augmented cognition* (pp. 449–458). Berlin, New York: Springer.

Belyavin, A. (2005). Construction of appropriate gauges for the control of Augmented Cognition systems. *Proceedings of the 1st International Conference on Augmented Cognition,* Las Vegas, NV, 22-27 July 2005.

Benson, A. J., Huddleston, J. H., & Rolfe, J. M. (1965). A psychophysiological study of compensatory tracking on a digital display. *Human Factors, 7*(5), 457-472.

Ben-Zadok, G., Hershkovitz, A., Mintz, E., & Nachmias, R. (2009). Examining online learning processes based on log files analysis: A case study. *Proceedings of the Fifth International Conference on Multimedia and ICT in Education, 1,* 55-59.

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., et al. (2004). Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset. *International Journal of Human-Computer Interaction 17*(2), 151–170.

Berka, C., Levendowski, D. J., Westbrook, P., Davis, G., & Lumicao, M. N. (2005). Implementation of a Closed-Loop Real-Time EEG-Based Drowsiness Detection System : Effects of Feedback Alarms on Performance in a Driving Simulator. *Proceedings of the 1st International Conference on Augmented Cognition,* Las Vegas, NV, 22-27 July 2005.

Berka, C., Levendowski, D. J., Davis, G., Whitmoyer, M., Hale, K., & Fuchs, S. (2006). Objective measures of situational awareness using neurophysiology technology. In D. D. Schmorrow, K. M. Stanney & L. M. Reeves (Eds.), *Foundations of Augmented Cognition* (pp. 145–154). Arlington and VA: Strategic Analysis, Inc.

Berka, C., Levendowski, D., Lumicao, M., Yau, A., Davis, G., Zivkovic, V., et. al. (2007). EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning and Memory Tasks. *Aviation Space and Environmental Medicine, 78*(5, Section II, Suppl.1), B231–B244.

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely Selective Attention: Eye-Tracking Studies of the Dynamic Allocation of Attention to Stimulus Features in Categorization. *Journal of Experimental Psychology, 35*(5), 1196–1206.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310.

Bland, J. M., & Altman, D. G. (1994). Correlation, regression, and repeated data. *British Medical Journal, 308(6933),* 896.

Bland, J. M., & Altman, D. G. (1995). Calculating correlation coefficients with repeated observations: Part 2 – correlation between subjects. *British Medical Journal, 310,(6980),* 633.

Bland, J. M., & Altman, D. G. (1996). Statistics Notes: Measurement error and correlation coefficients. *British Medical Journal*, 313(7048), 41–42.

Böcker, M. (2014). Nahinfrarotspektroskopie, funktionelle (fNIRS). In M. A. Wirtz (Ed.), *Dorsch – Lexikon der Psychologie* (18th ed., pp. 1077). Bern: Hogrefe Verlag.

Boiten, F. A., Frijda, N. H., & Wientjes, C. J. (1994). Emotions and respiratory patterns: review and critical analysis. *International journal of psychophysiology, 17*(2), 103-128.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.

Bosse, T., Memon, Z. A., & Treur, J. (2008). Adaptive Estimation of Emotion Generation for an Ambient Agent Model. In E. Aarts, J. L. Crowley, B. de Ruyter, H. Gerhauser, A. Pflaum, J. Schmidt et al., (Eds.), *AmI 2008,* LNCS, vol. 5355 (pp. 141–156). Heidelberg: Springer.

Bosse, T., Memon, Z. A., & Treur, J. (2009). An adaptive agent model for emotion reading by mirroring body states and Hebbian learning. In J.-J. Yang, M. Yokoo, T. Ito, Z. Jin & P. Scerri (Eds.*), Proceedings of the 12th international conference on principles of practice in multi-agent systems, PRIMA'09.* Lecture notes in artificial intelligence (Vol. 5925, pp. 552–562). Berlin, Heidelberg: Springer.

Bosse, T., Lambalgen, R. van, Maanen, P. P. van, & Treur, J. (2009). Attention Manipulation for Naval Tactical Picture Compilation. In R. Baeza-Yates, J. Lang, S. Mitra, S. Parsons & G. Pasi, (Eds.), *Proceedings of the 9th IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT'09* (Vol. 2, 450-457). Washtington, DC: IEEE Computer Society Press.

Boucsein, W., & Backs, R. W. (2000). Engineering psychophysiology as a discipline: historical and theoretical aspects. In: R.W. Backs & W. Boucsein (Eds.), *Engineering Psychophysiology. Issues and Applications (*pp. 3–30). Mahwah, N.J.: Lawrence Erlbaum Associates.

Bower, G. H. (1981). Mood and memory. American Psychologist, 36, 129-148.

Bradley, M. M., & Lang. P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry 25(1),* 49-59.

Bradley, M. M., & Lang, P. J. (2000). Affective reactions to acoustic stimuli. Psychophysiology, 37(2), 204–215.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation, *Psychophysiology, 45(4),* 602-607.

Brehm, J. W., & Self, E. (1989). The intensity of motivation. Annual Review of Psychology, 40(1), 109-131.

Broadbent, D. (1958). *Perception and Communications.* New York: Permagon Press.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42(3),* 361–377.

Brouwer, A. M., Zander, T. O., van Erp, J. B., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in neuroscience, 9,* 136.

Bruder, A. & Schwarz, J. (2019). Evaluation of diagnostic rules for real-time assessment of mental workload within a dynamic adaptation framework. Paper to be presented at the conference HCII 2009, 26.-31.07.19, Orlando FL, USA.

Buchanan, D., & Huczynski, A. (1997). *Organizational Behaviour: An Introductory Text (3rd ed.)*. London: Prentice Hall.

Bundele, M. M., & Banerjee, R. (2009). Detection of fatigue of vehicular driver using skin conductance and oximetry pulse: a neural network approach. *Proceedings of the 11th International Conference on Information Integration and web-based applications & services* (pp. 739-744). New York: ACM.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, et. al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. *Proceedings of the ACM Sixth International Conference on Multimodal Interfaces, ICMI 2004*, (pp. 205-211). New York: ACM.

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In A. Mital (Ed.), *Advances in Industrial Ergonomics and Safety (*481–485). London: Taylor & Francis.

Byrne, E., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological psychology, 42*(3), 249–268.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of emotions, 2*, 173-191.

Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. *European journal of applied physiology, 89*(3-4), 319-325.

Cain, B. (2007). *A review of the mental workload literature.* Toronto: Defence Research And Development.

Caldwell, J. A, Mallis, M. M., Caldwell, J. L., Paul, M. A., Miller, J. C., & Neri, D. F. (2009). Fatigue countermeasures in aviation. *Aviation, Space and Environmental Medicine, 80*(1), 29-59.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing, 1*(1), 18-37.

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology, 39*(1-4), 106-124.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is approach-related affect: Evidence and implications. *Psychological Bulletin, 135*(2), 183–204.

Cetintas, S., Si, L., Xin, Y. P., Hord, C., & Zhang, D. (2009). Learning to identify students' off-task behavior in intelligent tutoring systems. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. Graesser (Eds.), *Building learning systems that care: Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED '09)* (pp. 701–703). Amsterdam, the Netherlands: IOS Press.

Chu, Y., & Rouse, W. B. (1979). Adaptive allocation of decision making responsibility between human and computer in multi-task situations. *Proceedings of IEEE Transactions on Systems, Man and Cybernetics, SMC-9*(12), 769-778.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates.

Coury, B. G., & Semmel, R. D. (1996). Supervisory control and the design of intelligent user interfaces. In R. Parasuraman & M. Mouloua, (Eds.): *Automation and Human Performance: Theory and Applications* (pp. 201-219). New Jersey: Lawrence Erlbaum Associates.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag., 18*(1), 32–80.

Cropanzano, R., James, K., & Citera, M. (1993). A goal hierarchy model of personality, motivation, and leadership. *Research in organizational behavior, 15*, 267-322.

Cummings, M. L., Gao, F., & Thornburg, K. M. (2016). Boredom in the workplace: a new look at an old problem. *Human factors, 58*(2), 279-300.

De Greef, T., & Arciszewski, H. (2009). Triggering adaptive automation in naval command and control. In S. Cong (Ed.), *Frontiers in adaptive control* (pp. 165–188). Vienna: I-Tech.

De Greef, T., Lafeber, H., Van Oostendorp, H., & Lindenberg, J. (2009). Eye Movement as Indicators of Mental Workload to Trigger Adaptive Automation. In D. D. Schmorrow, I.V. Estabrooke & M. Grootjen (Eds.): *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, (pp. 219-228). Heidelberg: Springer.

De Lemos, J., Sadeghnia, G. R., Olafsdottir, I., & Jensen, O. (2008). Measuring Emotions Using Eye Tracking. In *Proc. Measuring Behavior 2008, 6th Int. Conf. Meth. & Tech. in Behavioral Research*, Maastricht, The Netherlands.

De Rivecourt, M., Kuperus, M. N., Post, W. J., & Mulder, L. J. M. (2008). Heart rate and eye movement measures as indices for mental effort during simulated flight. *Ergonomics, 51*(9), 1295–1319.

De Vicente, A., & Pain, H. (2002). Informing the detection of the students' motivational state: An empirical study. In: S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems* (pp. 933-943). Berlin, Heidelberg: Springer.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *BMJ, 329*(7458), 168-169.

Derbali, L., & Frasson, C. (2010). Prediction of Players Motivational States Using Electrophysiological Measures during Serious Game Play. *Conference on Advanced Learning Technologies, IEEE International*, Sousse, Tunisia, 498-502.

Dhillon, B. S. (2014). Human Error in Power Plant Maintenance. In B. S. Dhillon (Ed.), *Human Reliability, Error, and Human Factors in Power Generation* (pp. 135-149). Cham: Springer.

Di Nocera, F., Terenzi, M., & Camilli, M. (2006). Another look at scanpath: Distance to nearest neighbour as a measure of mental workload. In D. de Waard, K. A. Brookhuis & A. Toffetti (Eds.), *Developments in human factors in transportation, design, and evaluation* (pp. 295–303). Maastricht, Netherlands: Shaker Publishing.

Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., et al. (2010). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, and Environmental Medicine, 81*(4), 413-417.

Dickman, S. J., & Meyer, D. E. (1988). Impulsivity and speed-accuracy tradeoffs in information processing. *Journal of personality and social psychology, 54*(2), 274-290.

Diener, H., & Oertel, K. (2006). Experimental approach to affective interaction in games. In: Z. Pan et al. (Eds.), *Edutainment 2006*, LNCS 3942, 507–518.

Diethe, T. (2005). The future of augmentation managers. In D. D. Schmorrow (Ed.), *Foundations of augmented cognition* (pp. 631–640). Mahwah, NJ: Erlbaum.

Diethe, T., Dickson, B. T., Schmorrow, D., & Raley, C. (2004). Toward an augmented cockpit. In D. A.Vicenzi, M. Mouloua & P. A. Hancock (Eds.), *Human performance, situation awareness and automation: Current research and trends* (Vol. 2, pp. 65–69).Mahwah, NJ: Erlbaum.

DIN EN 10 075-1 (2000). *Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung. Teil 1: Allgemeines und Begriffe*. Berlin: Beuth.

Dinges, D. F., & Grace, R. (1998). *PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance.* Federal Highway Administration, Office of Motor Carriers (Rep. No. FHWA-MCRT-98-006).

Dirican, C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science, 3*, 1361-1367.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor : A gaze-reactive intelligent tutoring system. *Journal of Human Computer Studies, 70*(5), 377–398. doi:10.1016/j.ijhcs.2012.01.004

Dominguez, C. (1994). Can SA be defined? *Situation awareness: Papers and annotated bibliography*, 5-15.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Auflage). Berlin, Heidelberg: Springer.

Dorneich, M. C., Passinger, B., Hamblin, C., Keinrath, C., Vasek, J., Whitlow, et al. (2011). The Crew Workload Manager: An Open-loop Adaptive System Design for Next Generation Flight Decks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 16–20.

Dorneich, M. C., Whitlow, S. D., Ververs, P. M., Mathan, S., Raj, A., Muth, E., et al. (2004). *DARPA improving warfighter information intake under stress—Augmented cognition concept validation experiment (CVE) analysis report for the Honeywell team (Contract No. DAAD16-03-C-0054).* Arlington, VA: Defense Advanced Research Projects Agency.

Duchowski, A. T. (2002). A breadth-first survey of eye tracking applications, *Behavior Research Methods, Instruments, and Computers 34*(4), 455-471.

Durso, F.T. & Dattel, A.R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application*. Aldershot, UK: Ashgate, 137-154.

Edwards, T. (2013). *Human Performance in Air Traffic Control*. Unpublished doctoral dissertation, University of Nottingham, Nottingham, United Kingdom.

Egelund, N. (1982). Spectral analysis of heart rate variability as an index of driver fatigue. *Ergonomics, 25*(7), 663-672.

Eid, M. (2014). Validität, konvergente. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (18. Aufl., S. 1611). Bern: Verlag Hogrefe Verlag.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1982). What emotion categories or dimensions can observers judge from facial behavior? In P. Ekman (Ed.), *Emotion in the human face* (pp. 39-55). New York: Cambridge University Press.

Elkin-Frankston, S., Bracken, B. K., Irvin, S., & Jenkins, M. (2017). Are Behavioral Measures Useful for Detecting Cognitive Workload During Human-Computer Interaction? In T. Ahram & W. Karwowski (Eds.), *Advances in Intelligent Systems and Computing,* Volume 494 (pp. 127–137). Cham: Springer.

Elliot, A. J., Eder, A. B., & Harmon-Jones, E. (2013). Approach–avoidance motivation and emotion: Convergence and divergence. *Emotion Review, 5*(3), 308-311.

Emotiv (2013). *User Manual for Software Development Kit Release 2.0.0.20*, Hong Kong: Emotiv Ltd.

Emotiv Nutzerforum (2010a). Beitrag von Administrator „gmac", 03.09.2010 03:38:59. Online verfügbar unter:

https://www.emotiv.com/forums/topic/What_emotions_can_be_measured_with_EPOC_/ letzter Zugriff: 14.06.17

Emotiv Nutzerforum (2010b). Beitrag von Administrator „gmac“, 09.06.2010, 03:18:25. Online verfügbar unter: https://www.emotiv.com/forums/topic/Affective_Suite_Data/ Letzter Zugriff: 14.06.2017

Emotiv Nutzerforum (2011). Beitrag von Administrator „gmac“, 07.12.2011 17:03 Online verfügbar unter: https://www.emotiv.com/forums/topic/affective_values_in_affective_suite/ Letzter Zugriff: 14.06.2017

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

Endsley, M. R. (1995). A taxonomy of situation awareness errors. In R. Fuller, N. Johnson & N. McDonald (Eds.), *Human Factors in Aviation Operations* (pp. 287-292). Aldershot, UK: Avebury.

Endsley, M. R. (1999). Situation awareness and human error: Designing to support human performance. *Proceedings of the High Consequence Systems Surety Conference.* Albuquerque, NM: Sandia National Laboratory.

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.

Eysenck, M. W. (1997). *Anxiety and Cognition: A unified theory*. Hove, UK: Psychology Press.

Fairclough, S. (07. Juni 2012). *Troubleshooting and Mind-Reading: Developing EEG-based interaction with commercial systems.* [Blog-Beitrag]. Verfügbar unter: http://physiologicalcomputing.org/2012/06/troubleshooting-and-mind-reading-developing-eeg-based-interaction-with-commercial-systems/ Letzter Aufruf: 26.02.2018.

Farmer, E., & Brownson, A. (2003). *Review of Workload Measurement, Analysis and Interpretation Methods.* Report prepared for EUROCONTROL INTEGRA programme (CARE-Integra-TRS-130-02-WP2). Verfügbar unter: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.3382&rep=rep1&type=pdf Letzter Aufruf: 12.03.2019.

Faulstich, M. E., Williamson, D. A., McKenzie, S. J., Duchmann, E. G., Hutchinson, K. M., & Blouin, D. C. (1986). Temporal stability of psychophysiological responding: A comparative analysis of mental and physical stressors. *International Journal of Neuroscience, 30*(1-2), 65-72.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a Characterization of Adaptive Systems: A Framework for Researchers and System Designers. *Human Factors, 54*(6), 1008–1024.

Fowles-Winkler, A. M. (2003). Modelling with the integrated performance modelling environment (IPME). In A. Verbraeck & V. Hlupic (Eds.), *Proceedings of the 15th European Simulation Symposium* (pp. 26–29). Erlangen, Germany: SCS Publishing House.

Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., & Beller, J. (2011). Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cognition, Authority & Work, 14*(1), 3-18.

Flin R., O'Connor, P., & Crichton, M. (2008). *Safety at the Sharp End: A Guide to Non-Technical Skills*. Aldershot, UK: Ashgate.

Fracker, M. L. (1991). *Measures of situation awareness: An experimental evaluation*. Technical Report AL-TR0191-0127. Wright-Patterson AFB OH: Armstrong Laboratory.

Frank, G. R. (2007). *Monitoring seated postural responses to assess cognitive state.* Unpublished master's thesis, University Of Pittsburgh, PA.

Freedy, A., Madni, A., & Samet, M. (1985). Adaptive user models: Methodology and applications in man-computer systems. In W. B. Rouse (Ed.), *Advances in manmachine systems research* (vol. 2, pp. 249-293). Greenwich, CT: JAI Press.

Freeman, F. G., Mikulka, P. I., Prinzel, L. I., & Scerbo, M. W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology 50*(1), 61-76.

Fuchs, S., Hale, K. S., Stanney, K. M., Berka, C., Levendowski, D., & Juhnke, J. (2006). Physiological Sensors Cannot Effectively Drive System Mitigation Alone. In D. D. Schmorrow, K. M. Stanney, & L. M. Reeves (Eds.), *Foundations of Augmented Cognition (2nd Ed.)* (pp. 193–200). Arlington and VA: Strategic Analysis, Inc.

Fuchs, S., Hale, K. S., Stanney, K. M., Juhnke, J., & Schmorrow, D. D. (2007). Enhancing Mitigation in Augmented Cognition. *Journal of Cognitive Engineering & Decision Making, 1*(3), 309–326.

Fuchs, S., Schwarz, J. (2014). Vom passiven Werkzeug zum sozialen Akteur: Ansatz einer ganzheitlicheren Betrachtung adaptiver automatisierter Systeme. In M. Grandt & S. Schmerwitz (Eds.), *Der Mensch zwischen Automatisierung, Kompetenz und Verantwortung (56. DGLR Fachausschusssitzung Anthropotechnik, Ottobrunn, 14.-15.10.2014, S. 285-288).* Bonn: Deutsche Gesellschaft für Luft- und Raumfahrt e.V.

Fuchs, S. & Schwarz, J. (2017). Towards a dynamic selection and configuration of adaptation strategies in Augmented Cognition. In D. D. Schmorrow and C. M. Fidopiastis (Eds.), *Augmented Cognition 2017, Part II, LNAI 10285* (pp. 1–15). Cham: Springer International Publishing AG.

Fuchs, S., Schwarz, J., & Werger, A. (2016). *Adaptive Mensch-Maschine-Interaktion: Ganzheitliche Onlinediagnose und Systemadaptierung*. Abschlussbericht zur Studie E/E4BX/EA192/CF215 (Dezember 2016). Wachtberg, Germany: Fraunhofer FKIE.

Gagnon, J. F., Tremblay, S., Lafond, D., Rivest, M., & Couderc, F. (2014). Sensor-Hub: A Real-Time Data Integration and Processing Nexus for Adaptive C2 Systems. *Proceedings of the 6th IARIA Adaptive Conference*, 63-67. Venice, Italy.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., & Dehais, F. (2015). Real-time state estimation in a flight simulator using fnirs. PLoS ONE 10:e0121279. doi: 10.1371/journal.pone.0121279

Gawron, V. J. (2016). Overview of Self-Reported Measures of Fatigue. *The International Journal of Aviation Psychology, 26*(3-4), 120-131.

Gawron, V. J., French, J., & Funke, D. (2001). An Overview of Fatigue. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, Workload and Fatigue* (pp. 581–595). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Ghergulescu, I., & Muntean, C. H. (2010). MoGAME: Motivation based Game Level Adaptation Mechanism. *Proceedings of the 10th Annual Irish Learning Technology Association Conference EdTech 2010*. Athlone, Ireland.

Ghergulescu, I. & Muntean, C. H. (2011). Learner motivation assessment with <e-adventure> game platform. *Proceedings of AACE E-LEARN-World Conference on E-Learning in Corporate, Government, Healthcare & Higher Education* (pp. 1212-1221). Honolulu, Hawaii.

Gimeno, T. P., Cerezuela, P. G., & Montanes, C. M. (2006). On the concept and measurement of driver drowsiness, fatigue and inattention: implications for countermeasures. *International journal of vehicle design, 42*(1-2), 67-86.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology, 56*(11), 1129-1135.

Goldberg, B. S., Sottilare, R. A., Brawner, K. W., & Holden, H. K. (2011). Predicting learner engagement during well-defined and ill-defined computer-based intercultural interactions. In: S. D'Mello, A. Graesser, B. Schuller & J.-C. Martin (Eds.), *International Conference on Affective Computing and Intelligent Interaction Part I, LNCS, vol. 6974* (pp. 538-547). Berlin, Heidelberg: Springer.

Grandt, M. (2004). *Erfassung und Bewertung der mentalen Beanspruchung mittels psychophysiologischer Messverfahren*. FKIE-Bericht Nr. 88. Wachtberg: Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie.

Grassmann, M., Vlemincx, E., von Leupoldt, A., & Van den Bergh, O. (2016). The role of respiratory measures to assess mental load in pilot selection. *Ergonomics, 59*(6), 745-753.

Greenberg, J., & Baron, R. A. (2008). *Behavior in Organizations (9 th ed)*. Upper Saddle River, NJ: Pearson Education.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112.

Gu, H., & Ji, Q. (2004). An automated face reader for fatigue detection. *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 111-116.

Haarmann, A., Boucsein, W., & Schaefer, F. (2009). Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. *Applied ergonomics*, *40*(6), 1026–1040.

Hadley, G. A., Prinzel, L. J., Freeman, F. G., & Mikulka, P. J. (1999). Behavioral, subjective and psychophysiological correlates of various schedules of short-cycle automation. In M. W. Scerbo & M. Mouloua (Eds.), *Automation Technology and Human Performance* (pp.139-143). Mahwah, NJ: Lawrence Erlbaum Assoc., Inc.

Hale, K. S., Fuchs, S., Axelsson, P., Berka, C., & Cowell, A. J. (2008). Using physiological measures to discriminate signal detection outcome during imagery analysis. *Proceedings of the Human Factors and Ergonomics Society 52th Annual Meeting*, 182–186.

Hancock, P. A., & Verwey, W. B. (1997). Fatigue, workload and adaptive driver systems. *Accident Analysis and Prevention, 29*(4), 495-506.

Hancock, P. A., & Chignell, M. H. (1987). Adaptive Control in Human-Machine Systems. In P. A. Hancock (Ed.), *Human Factors Psychology* (pp. 305-345). Amsterdam: North-Holland.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE transactions on Systems, Man, and Cybernetics, 18*(4), 647-658.

Hancock, P. A., & Scallen, S. F. (1998). Allocating functions in human-machine systems. In R. R Hoffman, M. F. Warm, & J. S Sherrick. (Eds.), *Viewing Psychology as a Whole: The Integrative Science of William N. Dember* (pp. 509–539). Washington, D. C.: American Psychological Association.

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, space, and environmental medicine, 69*(4), 360-367.

Hanley, J. A., & McNeil, B. J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology, 143*(1), 29-36.

Hargutt, V. (2003). *Das Lidschlagverhalten als Indikator für Aufmerksamkeits- und Müdigkeitsprozesse bei Arbeitshandlungen*. Düsseldorf: VDI Verlag.

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.

Hartel, C., Smith, K., & Prince, C. (1991). Defining aircrew situation awareness: Searching for mishaps with meaning. In D. Jensen (Ed.) *Proceedings of the 6th International Symposium on Aviation Psychology.* Columbus, Ohio: OSU.

Hartley, L., Horberry, T., Mabbott, N., & Krueger, G. P. (2000). *Review of Fatigue Detection and Prediction Technologies.* Melbourne, Australia: National Road Transport Commission.

Heller, O. (1982). *Theorie und Praxis des Verfahrens der Kategorienunterteilung (KU).* Würzburg: Würzburger Psychologisches Institut.

Hershkovitz, A., & Nachmias, R. (2011). Online persistence in higher education web-supported courses. *The Internet and Higher Education, 14*(2), 98-106.

Hilburn, B. J., Byrne, E., & Parasuraman, R. (1997). The effect of adaptive air traffic control (ATC) decision aiding on controller mental workload. In M. Mouloua & J. M. Koonce (Eds.), *Human-automation interaction: Research and practice* (pp. 84-91). Mahwah, NJ: LEA.

Hilburn, B. & Jorna, P. (2001). Workload and air traffic control. In P. A. Hancock and P. A. Desmond (Eds.), *Stress, workload and fatigue: Theory, research and practice*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Hincks, S. W., Afergan, D., & Jacob, R. J. K. (2016). Using fNIRS for Real-Time Cognitive Workload Assessment. In D. Schmorrow & C. Fidopiastis (Eds), *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience.* AC 2016. Lecture Notes in Computer Science, vol 9743 (pp. 198-208). Cham: Springer.

Hirshfield, L., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E., Jacob, R., et al. (2009). Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload. In D. D. Schmorrow, I. V. Estabrooke, & M. Grootjen (Eds.), *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, Number 5638 in Lecture Notes in Computer Science (pp. 239-247). Berlin, Heidelberg: Springer.

Hockey, G. R. J. (1970). Effect of loud noise on attentional selectivity. *Quarterly Journal of Experimental Psychology, 22*(1), pp. 28-36.

Hockey, G. R. J. (1986). Changes in operator efficiency as function of effects of environmental stress, fatigue and circadian rhythm. In K. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. 2. Cognitive processes and performance* (pp. 1-49). NY: John Wiley and Sons.

Hockey, G. R. J. (2003). Operator functional state as a framework for the assessment of performance degradation. In G. R. J. Hockey, A. W. K. Gaillard & O. Burov (Eds.), *Operator Functional State* (pp.8–23). Amsterdam: IOS Press.

Hockey, R. (2013). *The psychology of fatigue: Work, effort and control*. Cambridge: Cambridge University Press.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W.C. (1973). Quantification of sleepiness: a new approach. *Psychophysiology,10*(4), 431–436.

Hogervorst, M. A., Brouwer, A. M., & van Erp, J. B. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience, 8*(322). doi: 10.3389/fnins.2014.00322.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hudlicka, E., & McNeese, D. (2002). Assessment of user affective and belief states for interface adaptation: Application to an Air Force pilot task. *User Modeling and User-Adapted Interaction, 12*(1), 1-47.

Hursh, S. R., Balkin, T. J., Miller, J. C., & Eddy, D. R. (2004). The fatigue avoidance scheduling tool: Modeling to minimize the effects of fatigue on cognitive performance. *SAE Transactions, 113*(1), 111–119.

Hurwitz, J. B., & Wheatley, D. J. (2002). Using driver performance measures to estimate workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*(22), 1804-1808.

IBM Corp. (2011). *IBM SPSS Statistics for Windows, Version 20.0.* Armonk, NY: IBM Corp.

Inagaki, T. (2003). Adaptive Automation: Sharing and Trading of Control, In. E. Hollnagel (Ed.) *Handbook of Cognitive Task Design,* (pp. 147-169). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human computer interaction. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems,* April 24-29, 2004, Vienna, Austria (pp. 1477-1480). New York: ACM.

James, W. (1890).*The principles of psychology* (Vol 2). New York: Holt.

James, W. (1884). What is an emotion? *Mind, 9*, 188-205.

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue, *Proceedings of IEEE Transactions on Vehicular Technology, 53*(4), 1052-1068.

Johnson L. C. & Naitoh, P. (1974). *The operational consequences of sleep deprivation and sleep deficit* (NATO AGARDograph No. 193). Paris, France: Advisory Group for Aerospace Research and Development.

Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., & Berka, C. (2011). Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology, 87*(2), 241–250.

Johnston, M. R., Carpenter, A. C., & Hale, K. (2011). Test-Retest Reliability of CogGauge: A Cognitive Assessment Tool for SpaceFlight. In *International Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 565-571). Berlin, Heidelberg: Springer.

Johnston, W. A., & Dark, V. J. (1986). Selective attention. *Annual Review of Psychology, 37*(1), 43-75.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis, J. Haviland (Eds.), *Handbook of emotion, 2.edition* (pp. 220–235). New York: Guilford.

Jones, D., Hale, K., Dechmerowski, S., & Fouad, H. (2012). Creating Adaptive Emotional Experience During VE Training. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)* (pp. 1-10). Orlando, FL (Dec 3-6).

Jurcak, V., Tsuzuki, D., & Dan, I. (2007). 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *NeuroImage, 34*(4), 1600-1611.

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*(2), 113-153.

Kaber, D. B., Prinzel, L. J., Wright, M. C., & Clamann, M. P. (2002). *Workload-Matched Adaptive Automation Support of Air Traffic Controller Information Processing Stages.* Technical Paper

NASA/TP-2002-211932. Hampton, Virginia: National Aeronautics and Space Administration, Langley Research Center.

Kaber, D. B., & Wright, M. C. (2003). Adaptive automation of stages of information processing and the relation to operator functional states. In G. R. J. Hockey (Ed.), *Operator Functional States*. Vol 355 (pp. 204-223). IOS Press: NATO Science Series Sub Series I Life and Behavioral Science.

Kardan, S., & Conati, C. (2012). Exploring Gaze Data for Determining User Learning with an Interactive Simulation. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User Modeling, Adaptation, and Personalization* (126–138). Berlin Heidelberg: Springer.

Kaster, A., Tappert, E., Ruckert, C., & Becker, R. (2010). *Gestaltung ergonomischer Benutzungsschnittstellen für Asymmetric Warfare (GeBAW)*. Abschlussbericht. Wachtberg: Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE.

Kecklund, G., Åkerstedt, T., Sandberg, D.. Wahde, M., Dukic, T., Anund, A., et al. (2007). *DROWSI - State of the art review of driver sleepiness*. IVSS project report.

Keilhacker, P. (2013). *Subjektive Bewertung von Lärmminderungsmaßnahmen mittels Kategorien- und Verhältnisskalierung* (Inaugural-Dissertation). Eichstätt: Katholische Universität Eichstätt-Ingolstadt.

King, L.A. (2009). Visual navigation patterns and cognitive load. In D. D. Schmorrow, I. V. Estabrooke & M. Grootjen (Eds.), *Foundations of augmented cognition. Neuroergonomics and operational neuroscience. FAC 2009. LNCS vol 5638* (pp. 254–259). Berlin, Heidelberg: Springer.

Kleiniginna, P. R., & Kleiniginna, A. M. (1981a). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*(4), 345–379.

Kleiniginna, P. R., & Kleiniginna, A. M. (1981b). A categorized list of motivation definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*(3), 263–291.

Kok, A. (1997). Event-related-potential (ERP) reflections of mental resources: A review and synthesis, *Biological Psychology*, *45*(1-3), 19-56.

Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. Damos (Ed.), *Multiple Task Performance* (pp. 279-328). London: Taylor and Francis.

Krueger, G. P. (1989). *Sustained Work, Fatigue, Sleep Loss and Performance: A Review of the Issues (Reprint)*. USAARL Report 89-22. Online verfügbar: http://www.usaarl.army.mil/techreports/89-22.pdf (letzter Abruf: 21.01.18)

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18–27.

Lal, S. K., & Craig, A. (2001). A critical review of the psychophysiology of driver fatigue. *Biological psychology*, *55*(3), 173-194.

Larue, G. S., Michael, R., & Rakotonirainy, A. (2011). Drivers' Inability to Assess Their Level of Alertness on Monotonous Highways. In *Proceedings of the 8th International conference on managing fatigue in transportation, resources and health*. Fremantle, WA.

Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York: McGraw-Hill.

Lazarus, R. S. (1993). From psychological stress to the emotions: A history of a changing outlook. *Annual Review of Psychology*, *44*(1), 1–21.

Lazarus, R. S., & Eriksen, C .W. (1952). Effects of failure stress upon skilled performance. *Journal of Experimental Psychology*, *43*(2), 100-105.

Lee, C. M., & Narayanan, S. S. (2005). Towards detecting emotions in spoken dialogs. *IEEE Trans. on Speech & Audio Processing*, *13*(2), 293-303.

Lee, J., & See, K. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors, 46*(1), 50-80.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Lin, C. T., Ko, L. W., Chung, I. F., Huang, T. Y., Chen, Y. C., Jung T. P., et al. (2006). Adaptive EEG-based alertness estimation system by using ICA-based fuzzy neural networks, *IEEE Transactions on Circuits and Systems I: Regular Papers, 53*(11), 2469 -2476.

Llorente, A. M., Amado, A. J., Voigt, R. G., Berretta, M. C., Fraley, J. K., Jensen, C. L., & Heird, W. C. (2001). Internal consistency, temporal stability, and reproducibility of individual index scores of the Test of Variables of Attention in children with attention-deficit/hyperactivity disorder. *Archives of Clinical Neuropsychology, 16(6)*, 535-546.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Lockheed, M. (2005). *DARPA improving warfighter information intake under stress— Augmented cognition Phase 3 concept validation experiment (CVE) analysis report for the Lockheed Martin ATL team* (Tech. Rep. submitted to DARPA/IPTO). Arlington, VA: Defense Advanced Research Projects Agency.

Macdonald, J. S., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Attention, Perception, & Psychophysics, 73*(6), 1780-1789.

Mack, A. & Rock, I. (1998). *Inattentional Blindness*. Cambridge, MA: MIT Press.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology, 1*(1), 6-21.

Mahlke, S., & Minge, M. (2006*). Emotions and EMG measures of facial muscles in interactive contexts*, Proceedings of CHI 2006, Montreal, Canada.

Makeig, S., & Inlow, M. (1993). Lapse in alertness: coherence of fluctuations in performance and EEG spectrum. *Electroencephalography and clinical neurophysiology, 86*(1), 23-35.

Manzey, D. (1998). Psychophysiologie mentaler Beanspruchung. In F. Rösler (Ed.), *Ergebnisse und Anwendungen der Psychophysiologie. Enzyklopädie der Psychologie, Themenbereich C, Serie 1, Band 5* (pp. 799-864). Göttingen: Hogrefe.

Manzey, D. (2008). Systemgestaltung und Automatisierung. In P. Badke-Schaub, G. Hofinger & K. Lauche (Eds.), *Human Factors. Psychologie sicheren Handelns in Risikobranchen* (pp. 333-352). Heidelberg: Springer.

Manzey, D., & Bahner, J. E. (2005). Vertrauen in Automation als Aspekt der Verlaesslichkeit von Mensch-Maschine-Systemen. In K. Karrer, B. Gauss & C. Steffens (Eds.), *Mensch-Maschine-Systemtechnik aus Forschung und Praxis* (pp. 93–109). Düsseldorf: Symposion.

Marañon, G. (1924). Contribution a 1'etude de Taction emotive de 1'adrenaline. *Revue Française d'Endocrinologie, 2,* 301–325.

Marshall, S. (2000). *Method and Apparatus for Eye Tracking and Monitoring Pupil Dilation to Evaluate Cognitive Activity, U.S. Patent 6,090,051*. Washington, DC: U.S. Patent and Trademark Office.

Marshall, S. (2002). The index of cognitive activity: Measuring cognitive workload. *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants, 2002* (pp. 7.5–7.9). New York: IEEE.

Maslow, A. H. (1943). A Theory of Human Motivation. *Psychological Review, 50(4)*, 370-96.

Mathan, S., Erdogmus, D., Huang, Y., et al. (2008). Rapid image analysis using neural signals. *Proceedings of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* (pp. 3309–3314). New York : ACM Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology, 6*(3), 175-183.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion, 23*(2), 209-237.

May, J. F. & Baldwin, C. L. (2009). Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transportation Research Part F: Traffic Psychology and Behaviour, 12*(3), 218-224.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta psychologica, 75*(1), 75-89.

McIntire, L., McKinley, R. A., Goodyear, C., Merritt, M., Griffin, K., McIntire, J., et al. (2011). *Eye Tracking: An Alternative Vigilance Detector*. Report No.AFRL-RH-WP-TR-2012-0022. Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

McKinley, R. A., McIntire, L. K., Schmidt, R., Repperger, D. W., & Caldwell, J. C. (2011). Evaluation of eye metrics as a detector of fatigue. *Human Factors, 53*(4), 403-414.

McLeod, P. (1977). A dual-task response modality effect: support for multi-processor models of attention. *Quarterly Journal of Experimental Psychology, 29*(4), 651-667.

McQuiggan, S. W., & Lester, J. C. (2006). Diagnosing Self-efficacy in Intelligent Tutoring Systems: An Empirical Study. In M. Ikeda, K. D. Ashley & T. W. Chan (Eds.), *Intelligent Tutoring Systems*. ITS 2006. Lecture Notes in Computer Science, vol 4053. Springer, Berlin, Heidelberg. DOI https://doi.org/10.1007

Meghanathan, R. N., van Leeuwen, C., & Nikolaev, A. R. (2014). Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Frontiers in Human Neuroscience, 8*, 1063. http://doi.org/10.3389/fnhum.2014.01063

Meddis, R. (1982). Cognitive dysfunction following loss of sleep. In A. Burton (Ed.), *The pathology and psychology of cognition* (pp. 225-252). London: Methuen.

Menke, L. E., Best, C., Funke, G. J., & Strang, A. J. (2015). Warfighter acceptance of future physiological monitoring and augmentation: A coalition study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 125-129. Los Angeles, CA: SAGE Publications.

Mieg, H. P. (2006). *Vigilanzentwicklung unter nCPAP-Therapie beim obstruktiven Schlafapnoesyndrom unter besonderer Berücksichtigung der zirkadianen Rhythmik.* Unpublished doctoral dissertation, Freie Universität Berlin, Berlin, Germany.

Miller, S. (2001). *Workload Measures. National Advanced Driving Simulator*. Oakland, IA: The University of Iowa.

Miller, N. L., Matsangas, P., & Shattuck, L. G. (2008). Fatigue and its effect on performance in military environments. In P. A. Hancock, & J. L. Szalma (Eds.). *Performance under stress* (pp. 231–250). Burlington, VT: Ashgate Publishing.

Miller, N. L., Shattuck, L. G., & Matsangas, P. (2011). Sleep and fatigue issues in continuous operations: a survey of U.S. Army officers. *Behavioral sleep medicine, 9*(1), 53–65. doi:10.1080/15402002.2011.533994

Mittal, A., Kumar, K., Dhamija, S., & Kaur, M. (2016). Head movement-based driver drowsiness detection: A review of state-of-art techniques. In *IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 903–908). New York: IEEE.

Molloy, K., Griffiths, T. D., Chait, M., & Lavie, N. (2015). Inattentional deafness: visual load leads to time-specific suppression of auditory evoked responses. *Journal of Neuroscience, 35*(49), 16046-16054.

Morris, N. M., & Rouse, W. B. (1986). *Adaptive aiding for human-computer control: Experimental studies of dynamic task allocation* (Tech. Report AAMRL-TR-86- 005). Wright-Patterson Air Force Base, OH: Armstrong Aerospace Medical Research Laboratory.

Morrison, J. G., Kobus, D. A., & Brown, C. M. (2006). *Volume I: DARPA improving warfighter information intake under stress—Augmented cognition. Phase II concept validation* (Tech. Rep. 1940). San Diego, CA: Pacific Science and Engineering Group.

Moruzzi, G., & Magoun, H. W. (1949). Brain stem reticular formation and activation of the EEG. Electroencephalogr. *Clin. Neurophysiol., 1*(1-4), 455-473.

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology, 34*(2-3), 205-236.

Mulder, L. J. M., Kruizinga, A., Stuiver, A., Vernema, I., & Hoogeboom, P. (2004). Monitoring cardiovascular state changes in a simulated ambulance dispatch task for use in adaptive automation. In D. de Waard, K.A. Brookhuis and C.M. Weikert (Eds.), *Human factors in design*. Maastricht: Shaker Publishing, pp. 161-175.

Mulder, B., Rusthoven, H., Kuperus, M., de Rivecourt, M., & de Waard, D. (2007). Short-term heart rate measures as indices of momentary changes in invested mental effort. In D. de Waard, G. R. J. Hockey, P. Nickel, & K. A. Brookhuis (Eds.), *Human Factors Issues in Complex System Performance* (pp. 101 - 116). Maastricht, the Netherlands: Shaker Publishing.

Mulder, B., de Waard, D., Hoogeboom, P., Quispel, L., & Stuiver, A. (2008). Using Physiological Measures For Task Adaptation: Towards a Companion. In J. H. D. M. Westerink, M. Ouwerkerk, T. J. M. Overbeek, W. F. Pasveer, & B. de Ruyter (Eds.), *Probing Experience: From Assessment of User Emotions and Behaviour to Development of Products* (pp. 221 - 234). Dordrecht, The Netherlands: Springer.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustical Society of America, 93*(2), pp. 1097-1108.

Neerincx, M. A. (2003). Cognitive task load design: Model, methods and examples. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (pp. 283–305). Mahwah, NJ: Lawrence Erlbaum Associates.

Nguyen, T., Ahn, S., Jang, H., Jun, S. C., & Kim, J. G. (2017). Utilization of a combined EEG/NIRS system to predict driver drowsiness. *Scientific Reports, 7*, 43933. http://doi.org/10.1038/srep43933

Nicholls, A. R., Polman, R. C. J., & Levy, A. R. (2012). A path analysis of stress appraisals, emotions, coping, and performance satisfaction among athletes. *Psychology of Sport & Exercise*, 13*(3)*, 263-270.

Noronha, H., Sol, R., & Vourvopoulos, A. (2013). Comparing the Levels of Frustration between an Eye-Tracker and a Mouse: A Pilot Study. In *Human Factors in Computing and Informatics* (pp. 107-121). Berlin, Heidelberg: Springer.

North, A. R., & Riley, V. A. (1989). W/INDEX: A Predictive Model of Operator Workload. In G. R. McMillan, D. Beevis, E. Salas, M. H. Strub, R. Sutton & L. van Breda (Eds.), *Applications of Human Performance Models to Systems Design* (pp. 81 - 89). New York: Plenum Press.

Nwe, T. L., Wei, F. S., & De Silva, L. C. (2001). Speech based emotion classification. In *TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology* (Vol. 1, pp. 297 -301). New York: IEEE.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33*(1), 17-33.

Oken, B. S., Salinsky, M. C., & Elsas, S. M. (2006). Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical neurophysiology, 117*(9), 1885-1901.

Olsen, A. (2012). *The Tobii I-VT fixation filter.* Tobii AB: Danderyd, Sweden, 20 March 2012.

Olsson, P. (2007). *Real-time and offline filters for eye tracking*. Unpublished master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

Or, C. K., & Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics, 7*(2), 83-94.

Orasanu, J. & Martin, L. (1998). Errors in aviation decision making: a factor in accidents and incidents. *Proceedings of the 2nd Workshop on Human Error, Safety, and System Development,* 100-107.

Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, 91*(9), 1370 –1390.

Papenmeier, F., & Huff, M. (2010). DynAOI: A tool for matching eye-movement data with dynamic areas of interest in animations and movies. *Behavior Research Methods, 42*(1), 179–187.

Parasuraman, R. (2003). Neuroergonomics: research and practice. *Theoretical Issues in Ergonomic Science, 4*(1-2), 5–20.

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems*. Progress Report NAWCADWAR-92033-60 under Contract No. N62269-90-0022-5931. Warminster, PA.

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors, 38*(4), 665-679.

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 of types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30*(3), 286–297.

Pelz, W. (2004). *Kompetent führen.* Wiesbaden: Gabler.

Pfendler, C., Pitrella, F. D., & Wiegand, D. (1995). Messung der Beanspruchung bei der Systembewertung. *Bericht Nr. 115. Forschungsinstitut für Anthropotechnik,* Wachtberg.

Philip, P., Sagaspe, P., Taillard, J., Valtat, C., Moore, N., Akerstedt, et al. (2005). Fatigue, sleepiness, and performance in simulated versus real driving conditions. *Sleep, 28*(12), 1511-1516.

Philip, P., Taillard, J., Moore, N., Delord, S., Valtat, C., Sagaspe, P., et al (2006). The effects of coffee and napping on nighttime highway driving. *Annals of Internal Medicine, 144*(11), 785–791.

Phillips, R. O. (2015). A review of definitions of fatigue – And a step towards a whole definition. *Transportation Research Part F 29,* 48-56.

Pilcher, J. J., Nadler, E., & Busch, C. (2002). Effects of hot and cold temperature exposure on performance: a meta-analytic review. *Ergonomics, 45*(10), 682–698.

Pomplun, M., Reingold, E. M., & Shen, J. (2001). Investigating the visual span in comparative search: The effects of task difficulty and divided attention, *Cognition, 81*(2), B57-B67.

Ponder, E., & Kennedy, W. P. (1927). On the act of blinking. *Quarterly Journal of Experimental Physiology, 18*(2), 89-110.

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology, 40*(1-2), 187-195.

Porter, L. W., & Lawler, E. E. (1968). *Management attitudes and performance*. Homewood, IL: Richard D. Irwin Company.

Prinzel III, L. J., Pope, A. T., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Prinzel, L. J. (2001). *Empirical Analysis of EEG and ERPs for Psychophysiological Adaptive Task Allocation. NASA/TM-2001-211016*. Hampton, VA: National Aeronautics and Space Administration.

Prinzel III, L. J., Parasuraman, R., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). *Three experiments examining the use of electroencephalogram, event-related potentials, and heart-rate variability for real-time human-centered adaptive automation design. NASA/TP-2003-212442*. Hampton, VA: National Aeronautics and Space Administration.

Pröll, M. (2012). *Using a low-cost gyro and eeg-based input device in interactive game design.* Unpublished master's thesis, University of Technology, Graz, Austria.

Qi, Y., Reynolds, C., & Picard, R. W. (2001). The Bayes Point Machine for computer-user frustrationdetection via pressuremouse. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces* (pp. 1–5). New York: ACM.

Qu, L., Wang, N., & Johnson, W. L. (2005). Using Learner Focus of attention to Detect Learner Motivation Factors. In: L. Ardissono, P. Brna & A. Mitrovic (Eds.), *User Modelling 2005* (pp. 70–73). Heidelberg: Springer.

Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1*(2), 178-183.

Rasmussen, J. (1983). Skills, Rules and Knowledge; Signals, Signs and Symbols and other Distinctions in Human Performance Models. *IEEE Transactions on Systems, Man and Cybernetics, SMC 13*(3), 257–266.

Rauch, N., Kaussner, A., Krüger, H. - P., Boverie, S., & Flemisch, F. (2009). The importance of driver state assessment within highly automated vehicles. Paper presented at the 16th ITS World Congress, Stockholm, Sweden, 21.-25. September 2009.

Reeves, D. L., Winter, K. P., Bleiberg, J., & Kane, R. L. (2007). ANAM® Genogram: Historical perspectives, description, and current endeavors. *Archives of Clinical Neuropsychology, 22*, 15-37.

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics, 54*(10), 932-942.

Reuderink, B., Nijholt, A., & Poel, M. (2009) Affective Pacman: A Frustrating Game for Brain-Computer Interface Experiments. In *Third International Conference on Intelligent Technologies for Interactive Entertainment* (pp. 221-227). Berlin, Heidelberg: Springer.

Reyner, L. A. & Horne, J. A. (2000). Early morning driver sleepiness: Effectiveness of 200 mg caffeine. *Psychophysiology, 37*(2), 251–256.

Rheinberg, F., Vollmeyer, R. & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen (Langversion, 2001). *Diagnostica, 2*, 57-66.

Ribback, S. (2003). *Psychophysiologische Untersuchung mentaler Beanspruchung in simulierten Mensch-Maschine-Interaktionen*. Unpublished doctoral dissertation, Universität Potsdam, Potsdam, Germany.

Richards, J. M., & Gross, J. J. (2000). Emotion regulation and memory: The cognitive costs of keeping one's cool. *Journal of Personality & Social Psychology, 79*(3), 410-424.

Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution, *Psychophysiology 22*(2), 204-207.

Roethlisberger, F. J., & Dickson, W.J. (1939). *Management and the Worker. An Account of a Research Program Conducted by the Western Electric Company*. Hawthorne Works, Chicago. Harvard University Press: Cambridge.

Rohmert, W. (1984). Das Belastungs-Beanspruchungskonzept. *Zeitschrift für Arbeitswissenschaft, 38*(4), 193–200.

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration?. *Biological Psychology, 34*(2-3), 259-287.

Roscoe, A. H. (1993). Heart rate as a psychophysiological measure for inflight workload assessment. *Ergonomics, 36*(9), 1055-1062.

Roth, W. T. (1983). A comparison of P300 and skin conductance response. In A. W. K. Gaillard & W. Ritter (Eds.), *Tutorials in event related potential research: Endogenous components* (pp. 177–199). Amsterdam: North-Holland Publishing.

Rouse, W. B. (1976). Adaptive allocation of decision making responsibility between supervisor and computer. In T. B. Sheridan & G. Johannsen (Eds.), *Monitoring behavior and supervisory control* (pp. 295-306). New York: Plenum.

Rouse, W. B. (1977). Human-computer interaction in multitask situations. *IEEE Transactions on Systems, Man, and Cybernetics, 7*(5), 384-392.

Rouse, W. B. (1981). Human-computer interaction in the control of dynamic systems. *ACM Computing Surveys (CSUR), 13*(1), 71-99.

Rouse, W. B. (1988). Adaptive Aiding for Human/Computer Control. *Human Factors*, 30(4), 431–443.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin, 100*(3), 349-363.

Ruiz-Padial, E., Sollers, J. J., Vila, J., & Thayer, J. F. (2003). The rhythm of the heart in the blink of an eye: Emotion-modulated startle magnitude covaries with heart rate variability. *Psychophysiology, 40*(2), 306-313.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies, *Psychological Bulletin, 115*(1), 102-141.

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation Awareness Measurement: A Review of Applicability for C4i Environments. *Applied Ergonomics, 37*(2), 225-238.

Sarter, N. (1991). The flight management system: Pilots' interaction with cockpit automation. *Proceedings of the International Conference on Systems, Man and Cybernetics* (pp.1307-1310). Boston, MA: IEEE.

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors, 37*(1), 5-19.

Scallen, S. F., & Hancock, P. A. (2001). Implementing Adaptive Function Allocation. *International Journal of Aviation Psychology, 11*(2), 197–221. Retrieved from http://www.leaonline.com/doi/abs/10.1207/S15327108IJAP1102_05

Scallen, S. F., Hancock, P. A., & Duley, J. A. (1995). Pilot performance and preference for short cycles of automation in adaptive function allocation. *Applied Ergonomics, 26*(6), 397-403.

Scerbo, M. W. (1996). Theoretical Perspectives on Adaptive Automation. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 37–63). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Scerbo, M. W. (2001). Stress, workload and boredom in vigilance: a problem and an answer. In Hancock, P. & Desmond, P. (Eds.), *Stress, workload & Fatigue* (pp. 267 - 278). Mahwah, New Jersey, USA: Erlbaum.

Scerbo, M. W., Freeman, F. G., & Mikulka, P. J. (2000). A biocybernetic system for adaptive automation. In R. Backs & W. Bouscein (Eds.), *Engineering psychophysiology: Issues and applications* (pp. 241–253). Mahwah, NJ: Erlbaum.

Schaaff, K., Degen, R., Adler, N., & Adam, M. T .P. (2012): Measuring affect using a standard mouse devise. *Biomedical Engineering, 57* (SUPPL.1 TRACK-N), 761-765. https://doi.org/10.1515/bmt-2012-4013

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379-399.

Schaub H. (2008). Wahrnehmung, Aufmerksamkeit und „Situation Awareness“ (SA). In P. Badke-Schaub, G. Hofinger & K. Lauche (Eds.), *Human Factors. Psychologie sicheren Handelns in Risikobranchen* (pp. 59-76). Berlin, Heidelberg: Springer.

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, *51*(7), 982-1010.

Schmidt, E. A. (2010). *Die objektive Erfassung von Müdigkeit während monotoner Tagfahrten und deren verbale Selbsteinschätzung durch den Fahrer*. Unpublished doctoral dissertation, Heinrich Heine Universität, Düsseldorf, Germany.

Schmidt, E. A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., & Kincses, W. E. (2009). Drivers’ misjudgement of vigilance state during prolonged monotonous daytime driving. *Accident Analysis and Prevention, 41*(5), 1087–1093.

Schmidtke, H. (1989). Wachsamkeit (Vigilanz). In H. J. Bullinger, H. W. Jürgens & W. Rohmert (Eds.), *Handbuch der Ergonomie, Band 2* (Kapitel 8.3.2), Koblenz: Bundesamt für Wehrtechnik und Beschaffung.

Schmidtke, H., & Micko, H. C. (1964). *Untersuchungen über die Reaktionszeit bei Dauerbeobachtungen.* Köln, Opladen: Westdeutscher Verlag.

Schmorrow, D. D. & Kruse, A. A. (2002). DARPA’s Augmented Cognition Program - Tomorrow’s Human Computer Interaction from Vision to Reality: Building Cognitively Aware Computational Systems. *Proceedings of IEEE Conference on Human Factors and Power Plants, 7*(1-4), doi: 10.1109/HFPP.2002.1042859.

Schoeffel, R. (1987). Zur Psychologischen Optimierung von Marine-Wachsystemen. In *Wehrpsychologische Untersuchungen, 22*(1). Bonn: Bundesministerium der Verteidigung.

Schwalm, M. (2009). *Pupillometrie als Methode zur Erfassung mentaler Beanspruchungen im automotiven Kontext.* Unpublished doctoral dissertation, Universität des Saarlandes, Saarbrücken, Germany.

Schwarz, J. (2013). Benutzerzustandserfassung zur Regelung Kognitiver Assistenz an Bord von Marineschiffen. In D. Söffker (Eds.), *2. Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten*. Duisburg-Essen: DuEPublico (Online-Publikation).

Schwarz, J., & Fuchs, S. (2014). Adaptive Automation als sozialer Akteur: Anforderungen an die Gestaltung aus psychologischer und systemtheoretischer Sicht. In M. Grandt & S. Schmerwitz (Eds.), *Der Mensch zwischen Automatisierung, Kompetenz und Verantwortung* (56. DGLR Fachausschusssitzung Anthropotechnik, Ottobrunn, 14.-15.10.2014, S. 107-123). Bonn: Deutsche Gesellschaft für Luft- und Raumfahrt e.V.

Schwarz, J., Fuchs, F., Becker, R., & Kaster, A. (2013). *Untersuchung der kognitiven Leistungsfähigkeit am Arbeitsplatz im Zwei-Wachen-Betrieb auf F125, sowie Ableitung von Maßnahmen zur Verbesserung der Durchhaltefähigkeit (UkLAZ)*. Abschlussbericht. Wachtberg: Fraunhofer FKIE.

Schwarz, J., & Witt, O. (2011). Evaluation of a touch-based user interface for a naval Command & Control System. *Proceedings of the 1st Conference on Maritime Human Factors, Ergoship 2011*, Gothenburg, Sweden, September 2011.

Schwarz, J., Bracco, F., Chiorri, C., Lommi, A., De Angelis, P. et al. (2012). FODAI - Fatigue and Overload Detection and Advising Interface. WP3 – Definition and Validation of the Methodology. D4.0 - Final Synthesis Report on Methodology Validation. CETENA, Genova (IT): Unclassified EDA Report.

Sciarini, L. W., & Nicholson, D. (2009). Assessing cognitive state with multiple physiological measures: A modular approach. In D. D. Schmorrow, I. V. Estabrooke & M. Grootjen (Eds.), *Augmented Cognition, HCII 2009*, LNAI 5638 (pp. 533-542). Berlin: Springer.

Selye, H. (1974). *Stress without Distress*. Philadelphia: Lippincott.

Shrout, P. E., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Sigari, M. H. (2009). Driver hypo-vigilance detection based on eyelid behavior. In *Proceedings of the Seventh International Conference on Advances in Pattern Recognition (ICAPR '09)* (pp. 426-429). New York, IEEE.

Silvagni, S., Napoletano, L., Graziani, I., LeBlaye, P., & Rognin, L. (2015). *Concept for Human Performance Envelope*. Future Sky Safety P6 Human Performance Envelope Report D6.1. Online verfügbar unter: https://www.futuresky-safety.eu/wp-content/uploads/2015/12/FSS_P6_DBL_D6.1-Concept-for-Human-Performance-Envelope_v2.0.pdf (Letzter Zugriff: 28.01.18).

Šimundić, A. M. (2008). Measures of diagnostic accuracy: basic definitions. *Medical and biological sciences, 22*(4), 61-65.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059-1074.

Singh, S. (2015). *Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey*. Traffic Safety Facts Crash, Stats. Report No. DOT HS 812 115. Washington, DC: National Highway Traffic Safety Administration.

Sirevaag, E. J., Rohrbaugh, J. W., Stern, J. A., Vedeniapin, A. B., Packingham, K. D., & LaJonchere, C. M. (1999). *Multi-dimensional characterization of operator state: a validation of oculumotor metrics*. Technical Report. Department of Psychology, St. Louis.

Sirevaag, E. J., & Stern, J. A. (2000). Ocular measures of fatigue and cognitive factors. In R.W. Backs & W. Boucsein (Eds.), *Engineering psychophysiology: Issues and applications* (pp. 269–287). Mahwah, NJ: Erlbaum.

Son, J., & Park, S. (2011). Cognitive workload estimation through lateral driving performance. In *Proceedings of the 16th Asia Pacific Automotive Engineering Conference, 2011*. Warrendale: PA: SAE Technical Paper.

Staal, M. A. (2004). *Stress, cognition, and human performance: A literature review and conceptual framework*. Hanover, MD: National Aeronautics & Space Administration.

Stanney, K. M., Schmorrow, D. D., Johnston, M., Fuchs, S., Jones, D., Hale, K.S., et al. (2009). Augmented Cognition: An Overview. In F.T. Durso (Ed.), *Reviews of Human Factors and Ergonomics* (Vol. 5, pp. 195–224). Santa Monica, CA: HFES.

Steinhauser, N. B., Pavlas, D., & Hancock, P. A. (2009). Design Principles for Adaptive Automation and Aiding. *Ergonomics in Design, 17*(2), 6–10.

Stefani, O., & Krüger, J. (2013). Chancen und Risiken zukunftsweisender Beleuchtungssysteme, *Zeitschrift für Arbeitswissenschaft, 67*(3), 175-179.

Steinhauser, N. B., Pavlas, D., & Hancock, P. A. (2009). Design Principles for Adaptive Automation and Aiding. *Ergonomics in Design, 17*(2), 6–10.

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: A possible measure of fatigue. *Human Factors, 36*(2), 285-297.

Stern, J. A., Boyer, D., Schroeder, D. J., Touchstone, R. M., & Stoliarov, N. (1996). *Blinks, saccades, and fixation pauses during vigilance task performance: II. Gender and time of day*. ADA307 024 FAA Office of Aviation Medicine—Civil Aeromedical Institute Publications, Aviation Medicine Reports. Washington, D.C.: Office of Aviation Medicine.

Strait, M., & Scheutz, M. (2014). What we can and cannot (yet) do with functional near infrared spectroscopy. *Frontiers in Neuroscience, 8*, 117. http://doi.org/10.3389/fnins.2014.00117

Stuiber, G. (2006). *Studie zur Untersuchung der Auswirkungen von Koffein auf den Pupillographischen Schläfrigkeitstest bei gesunden Probanden*. Unpublished doctoral dissertation, Universität Tübingen, Tübingen, Germany. Online verfügbar unter: http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-25246 (letzter Zugriff: 09.12.13)

Stuiver, A., Mulder, L. J. M., Brookhuis, K. A., de Waard, D., & Dijksterhuis, C. (2010). Adaptive task support based on dynamic human state estimation. In *Proceedings of 4th IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop (SASOW)* (pp. 153-158). New York: IEEE.

Swets, J. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Lawrence Erlbaum: Hilsdale, NJ.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). New York: Allyn and Bacon.

Tanaka, Y., & Yamaoka, K. (1993). Blink activity and task difficulty. *Perceptual and Motor Skills, 77*(1), 55-66.

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-Advisory Group for Aerospace Research and Development.

Thiffault, P., & Bergeron, J. (2003). Monotony of road environment and driver fatigue: A simulator study. *Accident Analysis and Prevention, 35*(3), 381–391.

Thorpy, M. J., & Billiard M. (2011). *Sleepiness: causes, consequences and treatment.* Cambridge: Cambridge University Press.

Timaeus, E., Lück, H. E., Klandt, H., & Schanderwitz, U. (1977). The PRS scale by Adair: An attempt to control motivations experimentally The PRS scale by Adair: An attempt to control motivations experimentally. *Zeitschrift für Experimentelle und Angewandte Psychologie, 24*(3), 510-518.

Ting, C. H., Mahfouf, M., Linkens, D. A., Nassef, A., Nickel, P., Roberts, A. C., et al. (2008). Real-time adaptive automation for performance enhancement of operators in a human-machine system. In *16th Mediterranean Conference on Control and Automation* (pp.552-557). New York: IEEE.

Tobii (2010). *Tobii Studio 2.2 User's Manual*. Stockholm, Sweden: Tobii Technology AB.

Tobii (2012). *Determining the Tobii I-VT Fixation Filter's Default Values* (White Paper). Tobii Technology AB, August 2012. Online verfügbar unter: https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/determining-the-tobii-pro-i-vt-fixation-filters-default-values.pdf. Letzter Zugriff: 09.02.18.

Tomarken, A. J. (1995). A psychometric perspective on psychophysiological measures. *Psychological Assessment, 7*(3), 387-395.

Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass, 8*(7), 328-341.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory. Vol. 3.* (pp. 200-239). Hillsdale, NJ: Erlbaum.

Trager, T. A. (1985). *Case study report on loss of safety system function events.* Report No. AEOD/C504. Washington, D.C: United States Nuclear Regulatory Commission.

Tremoulet, P., Barton, J., & Craven, P. (2005). *DARPA Improving Warfighter Information Intake Under Stress - Augmented Cognition Phase 3*. Concept Validation Experiment (CVE) Analysis Report for the Lockheed Martin ATL Team prepared under contract NBCHC030032. Arlington, VA: DARPA/IPTO.

Trimmel, M. & Poelzl, G. (2006). Impact of background noise on reaction time and brain DC potential changes of VDT-based spatial attention. *Ergonomics, 49*(2), 202-208.

Tupak S., Dresler T., Guhn A., Ehlis A., Fallgatter A., Pauli P., et al. (2014). Implicit emotion regulation in the presence of threat: neural and autonomic correlates. *Neuroimage 85*, 372–379.

Uhlarik, J., & Comerford, D. A. (2002). *A review of situation awareness literature relevant to pilot surveillance functions.* Collingdale, PA: Diane Publishing.

Uhlig, S. (2018). Heart Rate Variability: What remains at the end of the day? Doctoral dissertation, TU Chemnitz. Online verfügbar unter: http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-233101 (Letzter Zugriff: 26.02.19).

Uitenbroek, D. G. (1997). *SISA Simple Interactive Statistical Analysis*. Online verfügbar unter: http://www.quantitativeskills.com/sisa/index.htm.Letzter Zugriff: 09.02.18.

Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 271–360). San Diego, CA: Academic Press.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods 49*(2), 653–673.

Van Orden, K. F., Jung, T. P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological psychology, 52*(3), 221-240.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. (2001). Eye Activity Correlates of Workload during a Visuospatial Memory Task, *Human Factors, 43*(1), 111-121.

Veltman, J. A. (2002). A comparative study of psychophysiological reactions during simulator and real flight. *International Journal of Aviation Psychology*, *12*(1), 33-48.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological psychology, 42*(3), 323-342

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*(5), 656-669.

Veltman, J. A., & Jansen, C. (2003). Differentiation of Mental Effort measures: Consequences for Adaptive Automation. In G. R. J. Hockey, A. W. K. Gaillard & O. Burov (Eds.), *Operator Functional State* (pp. 249-259). Amsterdam: IOS Press.

Veltman, J. A. & Jansen, C. (2004). The adaptive operator. In D.A. Vincenzi, M. Mouloua & P.A. Hancock: *Human Performance, Situation Awareness and Automation Technology* (Vol. 2, pp. 7-10). Mahwah, NJ: Lawrence Erlbaum Associates.

Veltman, J. A., & Jansen, C. (2006). *The role of operator state assessment in adaptive automation*. TNO Report TNO-DV3 2005 A245. Soesterberg, Netherlands: TNO Report.

Veltman, H., Hockey, B., Schlegel, R. E., Fraser, W., & Burov, A. (2004). Introduction. In NATO Research & Technology Organisation (Ed.): *Operator Functional State Assessment.* NATO-RTO-TR-HFM-104. Neuilly-sur-Seine: NATO-RTO.

Veltman, H., & Vos, W. (2005). Facial Temperature as a Measure of Operator State. In D. D. Schmorrow (Ed.), *Foundations of Augmented Cognition* (pp. 293-301). Mahwah, NJ: Lawrence Erlbaum Associates.

Ververs, P. M., Whitlow, S. D., Doneich, M. C., & Mathan, S. (2005). Building Honeywell's Adaptive System for the Augmented Cognition Program. In D. D. Schmorrow (Ed.), *Foundations of Augmented Cognition* (pp. 460–468). Mahwah, NJ: Lawrence Erlbaum Associates.

Vicente, J., Laguna, P., Bartra, A., & Bailón, R. (2016). Drowsiness detection using heart rate variability. *Medical & biological engineering & computing*, *54*(6), 927-937.

Vogt, J., Hagemann, T., Kastner, M. (2006). The impact of workload on heart rate and blood pressure in en-route and tower air traffic control. *Journal of Psychophysiology 20*(4), 297–314.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

Wang, L., Duffy, V. G., & Du, Y. (2007). A Composite Measure for the Evaluation of Mental Workload. In: V.G. Duffy (Ed.). *Digital Human Modeling, HCII 2007, LNCS 4561* (pp. 460–466).

Waterhouse, I. K., & Child, I. L. (1953). Frustration and the quality of performance. *Journal of personality*, *21*(3), 298-311.

Weiner, E. A., & Concepcion, P. (1975). Effects of affective stimuli mode on eye-blink rate and anxiety. *Journal of clinical psychology*, *31*(2), 256-259.

Werger, A. (2016). *Entwicklung eines Regeleditors für die regelbasierte Interpretation von heterogenen Daten zur Leistungs- und Nutzerzustandsdiagnose*. Unpublished bachelor's thesis, Hochschule Bonn-Rhein-Sieg, Siegburg, Germany.

Whitlow, S., & Hayes, C. C. (2012). Considering Etiquette in the Design of an Adaptive System. *Journal of Cognitive Engineering and Decision Making*, *6*(2), 243–265.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67-85.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239-257). Hillsdale, NJ: Lawrence Erlbaum.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). New York: Academic Press.

Wickens, C. D. (1996a). Designing for Stress. In E. Salas, & J. E. Driskell (Eds.), *Stress & Human Performance* (pp. 279-295). Mahwah, NJ: Lawrence Erlbaum.

Wickens, C. D. (1996b). Situation awareness: impact of automation and display technology. In *NATO AGARD Aerospace Medical Panel Symposium on Situation Awareness*. Neuilly-Sur-Seine, France: AGARD CP-575.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159–177.

Wickens, C. D. (2005). Attentional tunneling and task management. *In Proceedings of the 13th International Symposium on Aviation Psychology* (pp. 620-625).

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *International Journal of Aviation Psychology, 19*(2), 182-199.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2016). *Engineering Psychology and Human Performance* (4th ed.). London, New York: Routledge.

Wickens, C. D., Gordon, S. E., & Liu, Y. (2004). *An Introduction to Human Factors Engineering* (2nd ed). Upper Saddle River, NJ: Pearson Prentice Hall.

Wickens, C. D., & Liu, Y. (1988). Codes and modalities in multiple resources: a success and a qualification. *Human Factors, 30*(5), 599-616.

Wickens, C. D., & McCarley, J. (2008). *Applied attention theory.* Boca-Raton, FL: Taylor & Francis.

Wiegmann, D. A., & Shappell, S. A. (2001). *A human error analysis of commercial aviation accidents using the human factors analysis and classification system (HFACS)*. Technischer Bericht (DOT/FAA/AM-01/3, FAA). Washington, D.C.: Federal Aviation Administration. Online verfügbar unter: http://www.dtic.mil/dtic/tr/fulltext/u2/a387808.pdf (letzter Zugriff: 09.02.18).

Wierwille, W. W., Ellsworth, L. A., Wreggit, S. S., Fairbanks, R. J., & Kirn, C. L. (1994). *Research on Vehicle-Based Driver Status/Performance Monitoring: Development, Validation, And Refinement of Algorithms For Detection Of Driver Drowsiness.* Washington, D.C.: National Highway Traffic Safety Administration.

Wientjes, C. J. E. (1992). Respiration in psychophysiology: methods and applications. *Biological psychology, 34(*2-3), 179–204.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin, 1*(6), 80-83.

Wilkinson, R. T., Edwards, R. S., & Haines, E. (1966). Performance following a night of reduced sleep. *Psychonomic Science, 5*(12), 471-472.

Wilson, G. F. (2000). Strategies for Psychophysiological Assessment of Situation Awareness. In M. R. Endsley, & D. J. Garland (Eds.), *Situation awareness*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, G. F. (2002). Adaptive aiding implemented by psychophysiologically determined operator functional state. *Proceedings of the NATO RTO-HFM symposium on the roles of humans in intelligent and automated systems*, Warsaw, 7–9 Oct 2002.

Wilson, G. F., Fraser, W., Beaumont, M., Grandt, M., Varoneckas, G., Veltman, H., et al. (2004). *Operator functional state assessment*. (NATO RTO Publication RTO-TR-HFM-104). Neuilly sur Seine: NATO Research and Technology Organization.

Wilson, G. F., & Russell, C. A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors*, *45*(3), 381–389. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14702990

Wilson, G. F., & Russell, C. A. (2006). Psychophysiologically Versus Task Determined Adaptive Aiding Accomplishment. In D. D. Schmorrow, K. M. Stanney & L. M. Reeves (Eds.),

Foundations of Augmented Cognition (pp. 201–207). Arlington and VA: Strategic Analysis, Inc.

Wirtz, M. & Nachtigall, C. (2002). *Deskriptive Statistik – Statistische Methoden für Psychologen, Teil 1*. Weinheim, München: Juventa.

Witt, O., Özyurt, E., Schwarz, J., Döring, B., & Dörfel, G. (2012). *Simulationsgestützte Entwicklung von Assistenzsystemen für Führungsaufgaben auf Marineschiffen unter Berücksichtigung des Demographischen Wandels (SADeWa)*. Abschlussbericht SADeWa. Wachtberg: Fraunhofer FKIE.

Wohleber, R.W., Matthews, G., Funke, G.J., Lin, J. (2016). Considerations in Physiological Metric Selection for Online Detection of Operator State: A Case Study. In D. D. Schmorrow, & C. Fidopiastis (Eds.), *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience* (pp. 428-439). Cham: Springer.

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition, 34*(3), 668–684.

Wood, J. G., & Hasset, J. (1983). Eyeblinking during problem solving. The effect of problem difficulty and internally vs. externally directed attention. *Psychophysiology, 20*(1), 18-20.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-Aware Tutors: Recognising and Responding to Student Affect. *International Journal of Learning Technology, 4*(3-4),129-164.

Wright, R. A. (1996). Brehm's theory of motivation as a model of effort and cardiovascular response. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The Psychology of Action: Linking Cognition and Motivation to Behavior* (pp. 424–453). New York: Guilford.

Wright, N. A., Stone, B. M., Horberry, T. J., & Reed, N. (2007). *A review of in-vehicle sleepiness detection devices.* Published project report 157. Wokingham, UK: TRL Limited.

Wright, F. P. (2010). *Emochat: Emotional instant messaging with the Epoc headset*. Unpublished master's thesis, University of Maryland, Baltimore County.

Yamada, F. (1998). Frontal midline theta rhythm and eyeblinking activity during a VDT task and a video game: useful tools for psychophysiology in ergonomics. *Ergonomics, 41*(5), 678-688.

Yamamoto, Y., & Isshiki, H. (1992). Instrument for controlling drowsiness using galvanic skin reflex. *Medical and Biological Engineering and Computing, 30*(5), 562-564.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459-482.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35.

Yuanyuan, P. (2006). *Development of EEG method for mental fatigue measurement*. Unpublished master's thesis, National University of Singapore, Singapore.

Zhang, Z., & Zhang, J. S. (2006). Driver fatigue detection based intelligent vehicle control. In *Proceedings of the 18th International Conference on Pattern Recognition,* (Vol. 2, pp.1262-1265). New York, IEEE.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. Delft: Delft University Press.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry, 39*(4), 561-577.

Anhang A. Versuchsmaterialien und ergänzende Auswertungen zu Experiment 1

A.1 Versuchsplan (Experiment 1)

		1. Termin				2. Termin			
VP	Faktor	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
1	Areas	2	5	2	5	2	5	2	5
	Kooperativität	100%	100%	70%	70%	100%	100%	70%	70%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
2	Areas	5	2	5	2	5	2	5	2
	Kooperativität	70%	100%	100%	70%	70%	100%	100%	70%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
3	Areas	2	5	2	5	2	5	2	5
	Kooperativität	70%	70%	100%	100%	70%	70%	100%	100%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
4	Areas	5	2	5	2	2	5	2	5
	Kooperativität	100%	70%	70%	100%	100%	70%	70%	100%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
5	Areas	2	2	5	5	2	2	5	5
	Kooperativität	100%	70%	100%	70%	100%	70%	100%	70%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
6	Areas	5	5	2	2	5	5	2	2
	Kooperativität	70%	100%	70%	100%	70%	100%	70%	100%
	Lärm	nein	nein	nein	nein	ja	ja	ja	ja
7	Areas	2	5	2	5	2	5	2	5
	Kooperativität	100%	100%	70%	70%	100%	100%	70%	70%
	Lärm	ja	ja	ja	ja	nein	nein	nein	nein
8	Areas	5	2	5	2	5	2	5	2
	Kooperativität	70%	100%	100%	70%	70%	100%	100%	70%
	Lärm	ja	ja	ja	ja	nein	nein	nein	nein
9	Areas	2	5	2	5	2	5	2	5
	Kooperativität	70%	70%	100%	100%	70%	70%	100%	100%
	Lärm	ja	ja	ja	ja	nein	nein	nein	nein
10	Areas	2	5	2	5	5	2	5	2
	Kooperativität	100%	70%	70%	100%	100%	70%	70%	100%
	Lärm	ja	ja	ja	ja	nein	nein	nein	nein
11	Areas	2	2	5	5	2	2	5	5
	Kooperativität	100%	70%	100%	70%	100%	70%	100%	70%
	Lärm	ja	ja	ja	ja	nein	nein	nein	nein
12	Areas	5	5	2	2	5	5	2	2
	Kooperativität	70%	100%	70%	100%	70%	100%	70%	100%
	Lärm	ja	ja	ja	ja	nein	nein	nein	nein

A.2 Einverständniserklärung (Experiment 1)

Fraunhofer FKIE

Einverständniserklärung

Sehr geehrte Versuchsteilnehmerin, sehr geehrter Versuchsteilnehmer,

vielen Dank, dass Sie sich bereit erklärt haben, an unserer Untersuchung teilzunehmen.

Bitte lesen Sie sich die folgenden Informationen aufmerksam durch. Wenn Sie noch Fragen haben, beantworten wir Ihnen diese gerne.

- Die Teilnahme an der Untersuchung ist freiwillig. Durch eine Verweigerung der Teilnahme werden Ihnen keine Nachteile entstehen. Die Teilnahme kann außerdem jederzeit ohne Nennung von Gründen abgebrochen werden, ohne dass Ihnen Nachteile entstehen.
- Die Untersuchung wird ausschließlich zu wissenschaftlichen Zwecken durchgeführt. Die aufgezeichneten Daten werden in keiner Weise zu einer Bewertung Ihrer Person verwendet.
- Die Erhebung und Datenauswertung erfolgt anonym. Für die Erhebung der Daten wird Ihnen eine Probandennummer zugeteilt, die bei allen Fragebögen und Tests angegeben wird.
- Die beabsichtigte Bekanntgabe des Untersuchungsergebnisses wird keine Einzeldaten enthalten und keinen Rückschluss auf Einzelpersonen zulassen.
- Die aufgezeichneten Daten (inkl. Videomaterial) sind nur den an der Untersuchung beteiligten Mitarbeitern des FKIE zugänglich.

Ich habe die vorliegenden Informationen zur Kenntnis genommen und wurde über Sinn und Zweck der Untersuchung informiert. Mit der Teilnahme an der Untersuchung und der Aufzeichnung meiner Daten unter den oben genannten Bedingungen bin ich einverstanden.

_____	_____	_____
Datum	Nachname, Vorname	Unterschrift

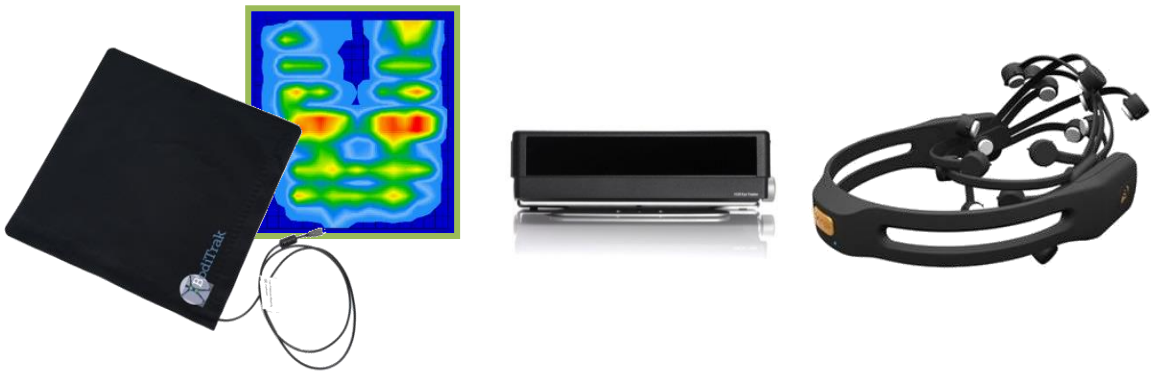
A.3 Instruktion (Experiment 1)

Liebe(r) Versuchsteilnehmer(in),

vielen Dank für Deine/ Ihre Bereitschaft, an unserem Versuch teilzunehmen!

Zweck der Untersuchung ist es, zu untersuchen, ob mit Hilfe verschiedener physiologischer Sensoren Zustandsveränderungen beim Menschen diagnostiziert werden können, die sich bei der Interaktion mit einem technischen System ergeben. Später soll diese Diagnosefunktion dazu genutzt werden, um die Funktionsweise und das Verhalten des technischen Systems (im Sinne eines adaptiven Systems) an den jeweiligen Nutzerzustand anzupassen.

Zur Diagnose des Nutzerzustands wollen wir einen Eyetracker einsetzen, der die Augenbewegungen erfasst, ein EEG, das die kortikale Aktivität misst und eine Drucksensormatte, die auf den Sitz gelegt wird und die Sitzposition bestimmt.



Der Ablauf der Untersuchung gliedert sich folgendermaßen:

1. Einweisung
2. Anpassen und Kalibrieren der physiologischen Sensoren
3. Fragebogen zum momentanen Befinden und zur Person
4. Vorabtests zur Rechenfähigkeit und zur Links-Rechts-Unterscheidung
5. Erfassung der Baseline der physiologischen Sensoren
6. Übungsszenario
7. Versuchsdurchführung (4 Szenarien à 10 Minuten)
8. Fragebogen nach jedem Szenario

Insgesamt wird die Untersuchungsdauer ca. 120 Minuten betragen.

Beschreibung der Experimentalaufgabe

Ziel der Experimentalaufgabe ist es, Luftkontakte auf einem simulierten Radarbildschirm zu steuern, indem über eine Eingabemaske Anweisungen an die Piloten gegeben werden. Ihr Rufname im Funkverkehr ist „Sunrise“.

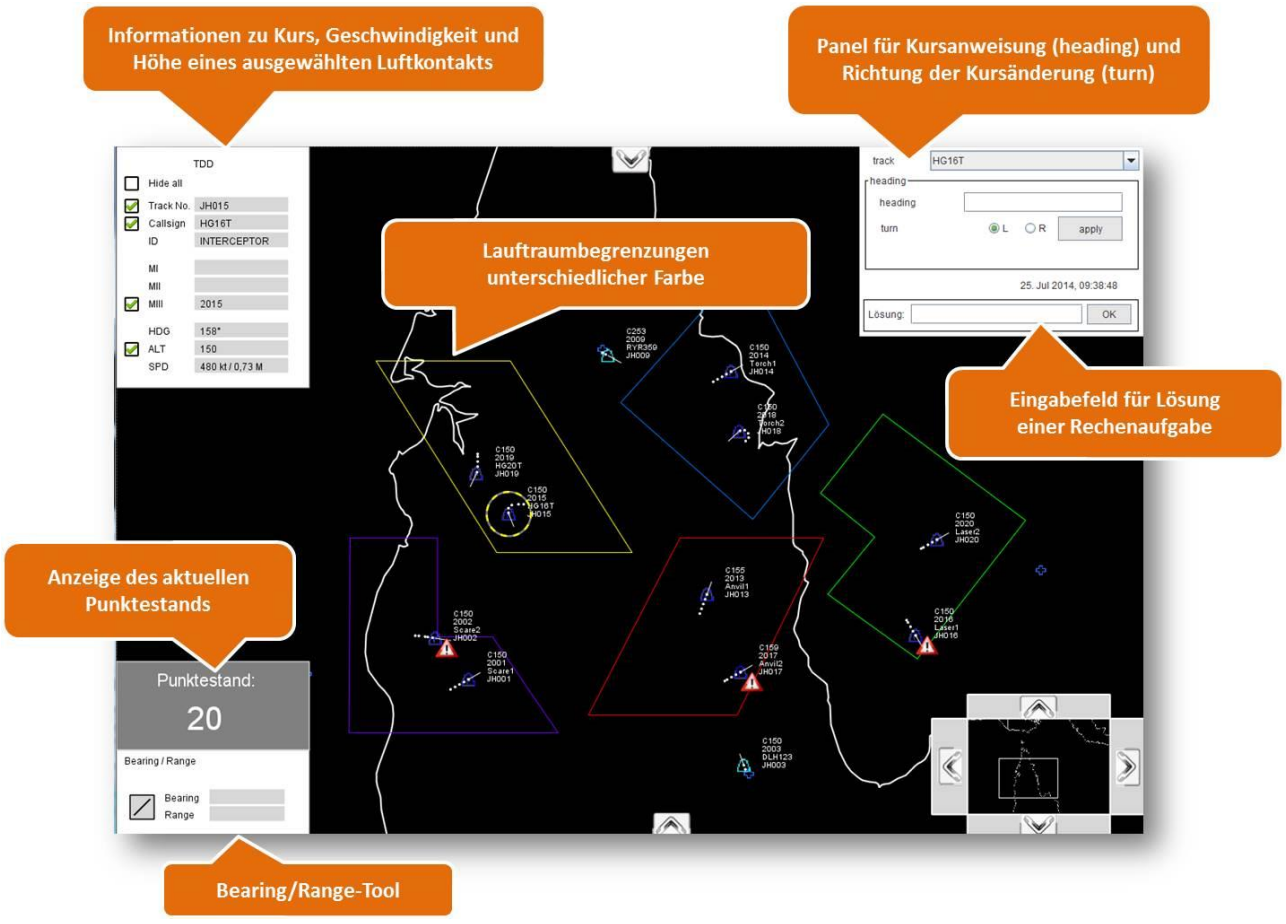
Solange Sie keine Richtungsanweisungen geben, wählen die Piloten ihre Flugrichtung selbständig. Dies kann dazu führen, dass sie die Grenzen ihres Luftraums verlassen oder mit anderen Luftkontakten kollidieren.

Aufgabe des Versuchsteilnehmers ist es, die Luftkontakte so anzuweisen, dass sie innerhalb ihrer Luftraumgrenzen (farblich gekennzeichnet) bleiben und Sicherheitsabstände zwischen den Luftkontakten eingehalten werden. Zu beachten ist, dass auch nicht steuerbarer Fremdverkehr die Areas durchfliegt, zu dem ebenfalls die Sicherheitsabstände eingehalten werden müssen.

Dabei gelten folgende Regeln:

- Sicherheitsabstand zwischen den Luftkontakten:
  - 5 nm horizontaler Abstand
  - 1000 ft vertikaler Abstand

- Abstand zur Luftraumgrenze:
  - 2 nm Abstand zwischen Kontakt und Luftraumgrenze



Punktevergabe

Wenn die Sicherheitsabstände für alle Kontakte eingehalten werden, werden jede 10 Sekunden 5 Pluspunkte gutgeschrieben.

Bei Abstandsverletzungen kommt es zu folgenden Punktabzügen:

- Abstandsverletzung zwischen Kontakten (Beinahe-Kollision): -20 Punkte
    - erneuter Punktabzug jeweils nach 30 Sekunden, wenn der Mindestabstand nicht wiederhergestellt ist

- Abstandsverletzung zur Luftraumgrenze oder Überschreiten der Luftraumgrenze: -10 Punkte
    - erneuter Punktabzug jeweils nach 60 Sekunden, wenn der Mindestabstand nicht eingehalten wird oder wenn sich der Kontakt noch außerhalb der Luftraumgrenze befindet.

Alarme

Alarme sollen als Warnung dienen, bei denen noch kein Punktabzug erfolgt. Die den Alarm auslösenden Luftkontakte werden zusätzlich mit einem roten Warndreieck gekennzeichnet. Der Alarm wird erst beendet, wenn für die betreffenden Kontakte eine Richtungsänderung angewiesen wurde. Das Warndreieck erlischt, wenn die Sicherheitsabstände wieder eingehalten werden.

Rechenaufgabe

Von Zeit zu Zeit werden über das Headset auf Englisch Rechenaufgaben gestellt. Angekündigt wird dies durch die Worte „Sunrise, calculate…“. Aufgabe ist es, die Lösung zu berechnen und in das dafür vorgesehene Formularfeld einzugeben. Dafür ist so lange Zeit, bis die nächste Rechenaufgabe gestellt wird.

Punktevergabe:

- Die richtige Lösung wird mit 5 Pluspunkten belohnt
- Bei keiner Lösung oder einer falschen Lösung gibt es 5 Punkte Abzug

Viel Spaß und viel Erfolg!

PS: Bitte andere Kollegen nicht über die Inhalte des Experiments informieren! Für uns ist es wichtig, dass die Versuchsteilnehmer die Versuchsbedingungen vorher nicht kennen. Danke!

A.4 Versuchsprotokoll (Experiment 1)

Versuchsprotokoll

Datum: _____ Probandennummer: _____

Uhrzeit: _____

Baseline Beginn:_____ Ende:_____

EEG Datei: _____ Body Tracker: _____

_____

_____

Übungsszenario Beginn:_____ Ende:_____

EEG Datei: _____ Body Tracker: _____

_____

_____

1. Durchgang Beginn:_____ Ende:_____

Bedingung: A: 2 ☐ 5 ☐ K: 100 ☐ 70 ☐ L: ja ☐ nein ☐

EEG Datei: _____ Body Tracker: _____

_____

_____

_____

2. Durchgang Beginn:_____ Ende:_____

Bedingung: A: 2 ☐ 5 ☐ K: 100 ☐ 70 ☐ L: ja ☐ nein ☐

EEG Datei: _____ Body Tracker: _____

_____

_____

_____

3. Durchgang Beginn:_____ Ende:_____

Bedingung: A: 2 ☐ 5 ☐ K: 100 ☐ 70 ☐ L: ja ☐ nein ☐

EEG Datei: _____ Body Tracker: _____

_____

_____

_____

4. Durchgang Beginn:_____ Ende:_____

Bedingung: A: 2 ☐ 5 ☐ K: 100 ☐ 70 ☐ L: ja ☐ nein ☐

EEG Datei: _____ Body Tracker: _____

_____

_____

_____

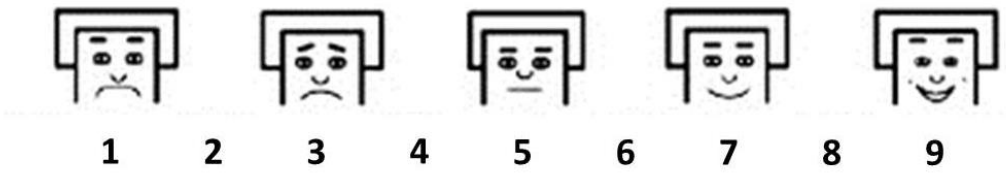
A.5 Fragebogen SAM - Self-Assessment-Manikin (Experiment 1)

Guten Tag,

im Folgenden möchten wir Sie bitten, Ihren Gefühlszustand während des vergangenen Versuchsdurchlaufs auf den drei Dimensionen Valenz (positiv vs. negativ), Arousal (ruhig vs. erregt) und Dominanz (dominierend vs. dominiert) zu bewerten.

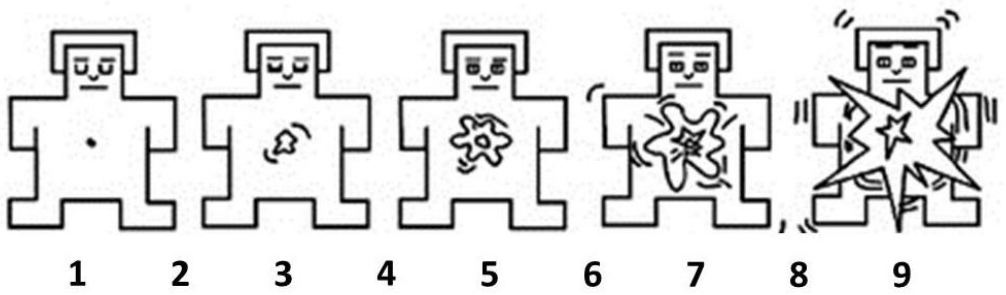
Hierzu werden so genannte Manikins (siehe Bild) abgebildet, die verschiedene Zustände symbolisieren. Die Manikins und die Zwischenstufen zwischen zwei Manikins sind durch Zahlen gekennzeichnet. Bitte wählen Sie auf der Ratingskala die Zahl eines Manikins oder einer Zwischenstufe aus, die am ehesten repräsentiert, wie Sie sich gefühlt haben.

Ihre Angaben werden selbstverständlich anonym behandelt und können nicht auf Ihre Person zurückgeführt werden.



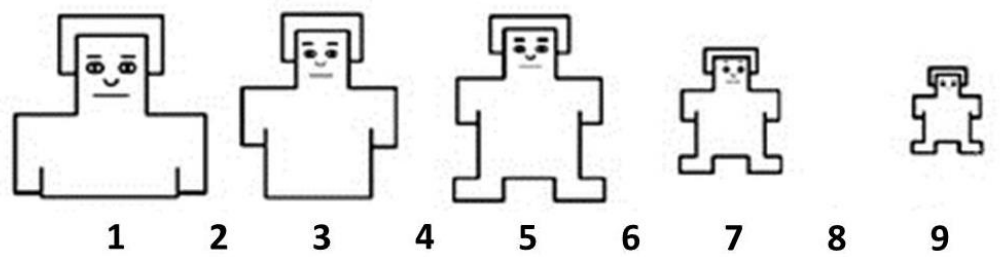
Bitte beurteilen Sie Ihren Gefühlszustand

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○



Bitte beurteilen Sie Ihren Erregungszustand

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○



Bitte beurteilen Sie Ihr Gefühl der Dominanz und Kontrolle

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○

A.6 Fragebogen zu individuellen Faktoren und Personenangaben (Experiment 1)

Müdigkeit

Bitte wählen Sie die Aussage aus, die Ihrem augenblicklichen Grad an Müdigkeit am ehesten entspricht.

- Fühle mich aktiv, vital, voll da, hellwach
- Habe einen klaren Kopf, bin aber nicht in Topform; kann mich konzentrieren
- Wach, aber entspannt; reagiere, bin aber nicht so ganz da
- Etwas benommen, schlaff
- Benommen, verliere das Interesse am Wachbleiben, tranig
- Kämpfe nicht mehr gegen den Schlaf, schlafe gleich ein; traumartige Gedanken

Motivation

Wie hoch ist Ihre Motivation, bei der Untersuchung eine gute Leistung zu erzielen

- sehr hoch
- hoch
- mittelmäßig
- gering
- sehr gering

Bitte geben Sie ihr Alter an: _____

Bitte geben Sie ihr Geschlecht an:

- Männlich
- Weiblich

Sind Sie Rechts- oder Linkshänder?

- Rechtshänder
- Linkshänder
- Beidhändig

Wie gut kennen Sie sich mit Computerspielen (Simulations- oder Strategiespiele) aus?

- gar nicht
- ein wenig
- gut
- sehr gut

A.7 Fragebogen zum Situationsbewusstsein und NASA-TLX (Experiment 1)

Situationsbewusstsein

Welche der folgenden Ereignisse sind im vergangenen Szenario aufgetreten?

- Boundary-Alarm (Annäherung an Luftraumgrenze)
- Kollisionsalarm
- Kontakt, der die Gebietsgrenze überschritten hat
- Kontakte, die kollidiert sind
- Rechenaufgabe, die richtig beantwortet wurde
- Rechenaufgabe, die falsch beantwortet wurde
- Rechenaufgabe, die ausgelassen wurde

Sonstige Ereignisse oder Auffälligkeiten (bitte im Kommentarfeld angeben).

NASA-TLX

Geistige Anforderungen

Wie beurteilen Sie den Schwierigkeitsgrad der Aufgaben in Bezug auf geistige Tätigkeiten und Wahrnehmungsvorgänge (z.B. Denken, Entscheiden, Erinnern, Beobachten, Suchen)?

sehr leicht			leicht			mäßig			schwierig			sehr schwierig		

Körperliche Anforderungen

Wie beurteilen Sie den Schwierigkeitsgrad der Aufgaben in Bezug auf Bedientätigkeiten?

sehr leicht			leicht			mäßig			schwierig			sehr schwierig		

Zeitliche Anforderungen

Wie stark empfanden Sie den Zeitdruck, der durch das Tempo, mit dem die Aufgabenelemente aufeinander folgten, verursacht wurde?

sehr gering			gering			mäßig			stark			sehr stark		

Leistung

Wie gut ist es Ihrer Ansicht nach gelungen, die gestellten Aufgaben zu erfüllen?

sehr gut			Gut			Mittel			schlecht			sehr schlecht		

Anstrengung

Wie sehr mussten Sie sich anstrengen, um die gestellten Aufgaben zu erfüllen?

sehr wenig			Wenig			mittel			stark			sehr stark		

Frustrationsniveau

Wie stark hat Sie die Bearbeitung der Aufgaben frustriert (z.B. verunsichert, entmutigt, irritiert, verärgert?

sehr wenig			Wenig			mittel			stark			sehr stark		

Hatten Sie Probleme mit bestimmten Aufgaben, wenn ja mit welchen und warum?

A.8 Abschlussfragen (mündlich) (Experiment 1)

VP-Nr._____

1. Wie war es für Dich/Sie, die verschiedenen Testdurchläufe durchzuführen?

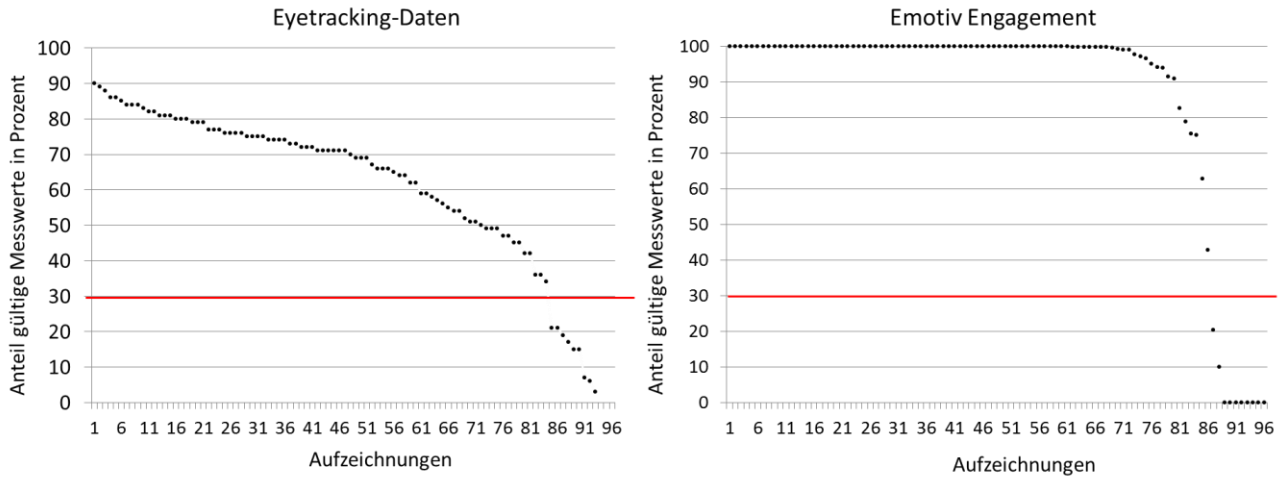
2. Hat sich Deine/Ihre Stimmung während des Versuchs geändert? Wenn ja, wann und woran lag das?

3. War es unangenehm, das EEG während des Versuchs auf zu haben?

4. Wenn ja, ab wann wurde es unangenehm und wie sehr hat es gestört?

5. Sonstige Anmerkungen

A.9 Aufzeichnungsqualität der Eyetracking-und EEG-Daten (Experiment 1)



Anmerkung: die rote Linie zeigt den Grenzwert von 30% an. Aufzeichnungen mit einem geringeren Anteil gültiger Messwerte wurden aus der Analyse ausgeschlossen.

Abbildung 51. Anteil gültiger Messwerte pro Aufzeichnung für die Eyetracking-Daten (links) und die Daten des Klassifikators Emotiv-Engagement (rechts) in absteigender Reihenfolge sortiert

A.10 Detailauswertung NASA-TLX (Experiment 1)

Tabelle 55. Ergebnisse der Varianzanalyse für den Gesamtscore und die Subskalen des NASA-TLX

	Haupteffekt Areas			Haupteffekt Kooperativität			Haupteffekt Lärm		
	$F(1,11)$	p	η_p^2	$F(1,11)$	p	η_p^2	F(1,11)	p	η_p^2
Gesamtscore	51.80	<.001	.83	20.26	.001	.65	.66	.44	.06
Geistige Anforder.	38.22	<.001	.78	9.96	<.01	.48	1.58	.24	.13
Körperl. Anforder.	12.85	<.01	.54	5.46	<.05	.33	1.92	.19	.15
Zeitl. Anforder.	69.01	<.001	.86	20.17	<.01	.55	.75	.41	.06
Leistung	49.69	<.001	.82	19.89	.001	.64	.70	.42	.06
Anstrengung	33.83	<.001	.76	12.34	<.01	.53	6.27	<.05	.36
Frustration	24.94	<.001	.69	11.56	<.01	.51	.17	.69	.02

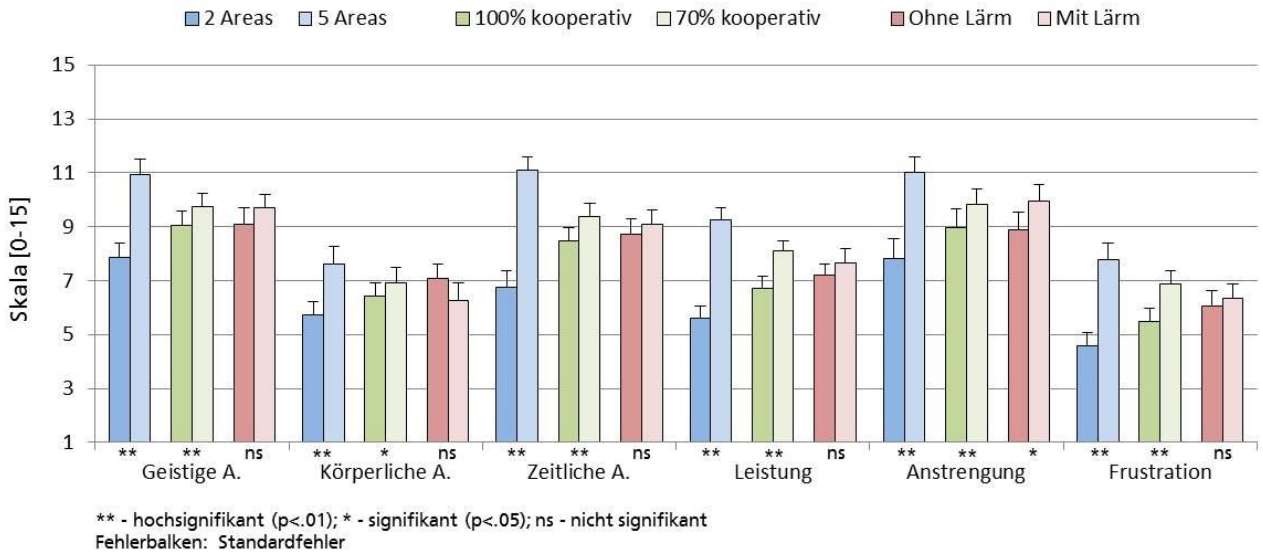


Abbildung 52. Unterschiede zwischen den Faktorstufen auf den Subskalen des NASA-TLX

A.11 Interaktionsdiagramme zur varianzanalytischen Auswertung in Abschnitt 4.3.1 (Experiment 1)

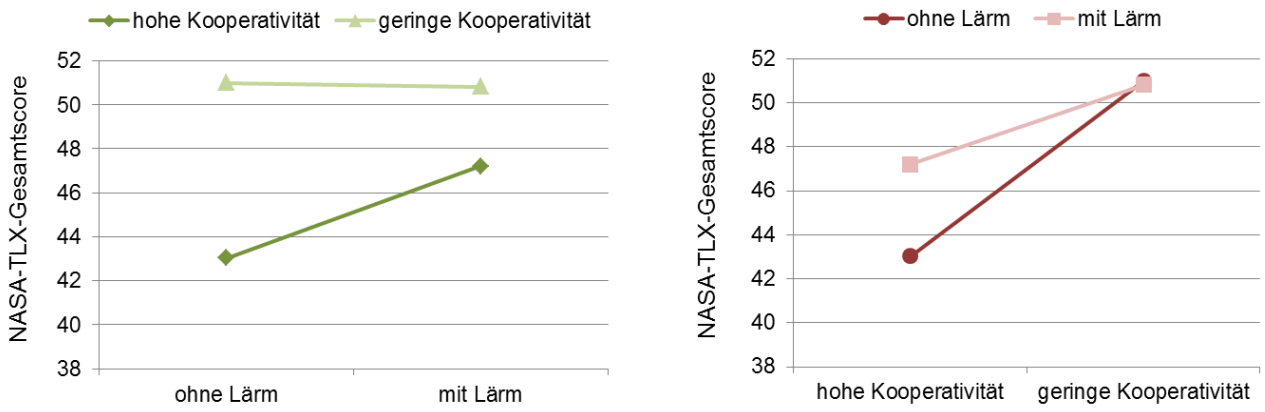


Abbildung 53. Interaktionsdiagramme zur Interaktion zwischen Kooperativität und Lärm hinsichtlich des NASA-TLX-Gesamtscore

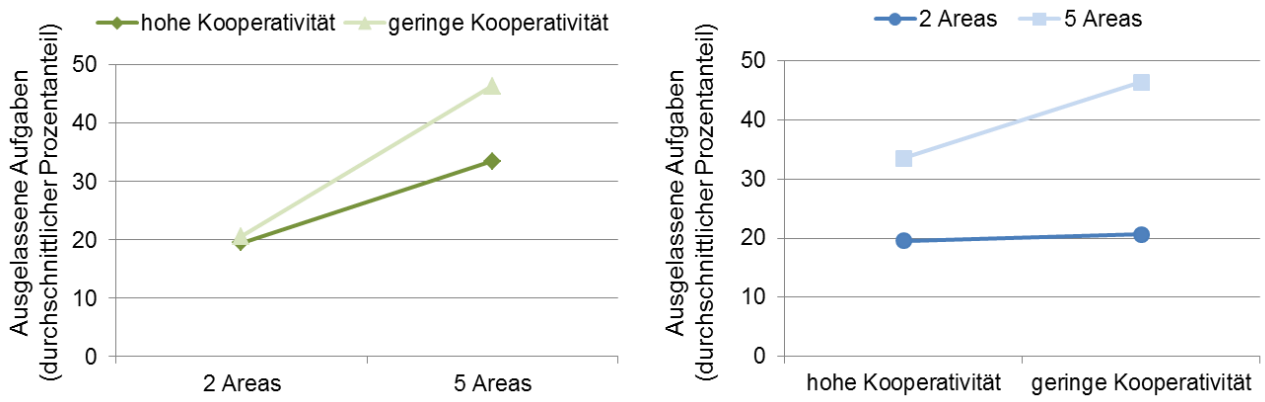


Abbildung 54. Interaktionsdiagramme zur Interaktion zwischen Anzahl Areas und Kooperativität hinsichtlich der Variable Ausgelassene Rechenaufgaben

Anhang B. Instruktion zu Experiment 2

Liebe(r) Versuchsteilnehmer(in),

vielen Dank für Deine/ Ihre erneute Bereitschaft, an unserem Versuch teilzunehmen!

Wie schon beim ersten Experiment möchten wir untersuchen, ob mit Hilfe verschiedener physiologischer Sensoren Zustandsveränderungen beim Menschen diagnostiziert werden können, die sich bei der Interaktion mit einem technischen System ergeben. Später soll diese Diagnosefunktion dazu genutzt werden, um die Funktionsweise und das Verhalten des technischen Systems (im Sinne eines adaptiven Systems) an den jeweiligen Nutzerzustand anzupassen.

Zusätzlich zu dem Eyetracker, dem EEG und der Sitzmatte, die bereits im ersten Experiment eingesetzt wurden, kommt nun noch der so genannte BioHarness hinzu. Dies ist ein Brustgurt (s. Abbildung), der verschiedene Vitalfunktionen, wie die Herzrate, die Atmungsrate, die Position des Körpers und die Körpertemperatur erfassen kann. Das Experiment dient einerseits dazu, die Diagnosefähigkeit dieses neuen Sensors zu testen, andererseits wollen wir auch untersuchen, inwiefern sich die Ergebnisse aus dem ersten Experiment replizieren lassen.

Der Ablauf der Untersuchung ist somit identisch mit der ersten Untersuchung. Er gliedert sich folgendermaßen:

1. Einweisung
2. Anpassen und Kalibrieren der physiologischen Sensoren
3. Fragebogen zum momentanen Befinden und zur Person
4. Vorabtests zur Rechenfähigkeit und zur Links-Rechts-Unterscheidung
5. Erfassung der Baseline der physiologischen Sensoren
6. Übungsszenario
7. Versuchsdurchführung (4 Szenarien à 10 Minuten)
8. Fragebogen nach jedem Szenario

Insgesamt wird die Untersuchungsdauer ca. 120 Minuten betragen.

Anhang C. Versuchsmaterialien und ergänzende Auswertungen zu Experiment 3

C.1 Aufgabenbeschreibung und ID-Kriterien (Experiment 3)

Description of the exercise

- The exercise consists of three scenarios of different workload levels (underload, normal, overload) and with a duration of 15 minutes each.

- The exercise contains the following tasks:
    1. Identification of new tracks on the TDA on the basis of ID-Crits (see below) or of tracks which show changes in their behaviour that call for a change of their ID
    2. Compilation of NRTT-Tracks on the basis of coordinates which are transmitted via headset or as a notification on the screen
    3. Warning of hostile/suspect/unknown tracks, which pass the safety range in direction to the own ship
    4. Engagement of hostile/suspect/unknown tracks, which are heading towards the own ship in spite of a warning and fall below a distance of 8nm to the own ship

    ➔ Priority of task execution (if more than one event occurs at the same time): 4 – 3 – 2 – 1
    ➔ Tasks have to be executed self-dependently

ID-CRITS

FRIEND/ASSUMED FRIEND:
- IFF Mode 4 positive reply
  OR:
- LINK (track number with five-digits)

NEUTRAL:
- IFF 3 + velocity < 400kts , altitude > 25.000ft
  OR:
- Following air route + velocity < 400kts, altitude > 25.000ft

SUSPECT/HOSTILE:
- No IFF Mode 4 + velocity > 400kts, altitude < 25.000ft,
  OR:
- No IFF Mode 4 + CPA < 1nm
  OR:
- No IFF Mode 4 + Split target

UNKNOWN: none of the criteria are valid

C.2 Beschreibung der Konsolenbedienung (Experiment 3)

Description of how to operate the console

The main screen of the demonstrator (see figure 1) is divided into several display areas. In the centre of the screen the TDA (Tactical Display Area) (1) is located, that shows the objects which have been detected by radar in the surrounding of the ship (in the following these objects are called “tracks”). The display areas around the TDA show further track information. For the tests in FODAI the following display areas will be used: The TAC (2), the Track List (3), the Warnings and Action options (4), the Effector Options (5) the Gun Control (6), located in the tab below the Threat evaluation and the panel Tasks (7), which was added to the main screen in order to provide information on NRT-Tracks in the test.

1. TDA

Table 1 shows a description of the different kinds of tracks on the TDA. Each track has a leader that represents the direction and the speed of the track.

Table 1: Explanation of the track symbols

Type of track	**symbol**
Unknown/ Pending	yellow symbols
Friend	blue symbols
Neutral	green symbols
Hostile	red symbols
Air contacts	the symbol is borderless at the bottom
Surface contacts	the symbol is completely framed
Subsurface contacts	the symbol is borderless at the top

Operations:

To hook a track: click on the track with the left mouse button. If a track is hooked information on the track is displayed in the TAC.

To select a track: Click on the ctrl+left mouse button or click ctrl and move the cursor to span a frame around one or more tracks which are to be selected

To create NRTT-Tracks click on the TDA with the left mouse button: select “create new track” in the pop-up-menu. The position, course and speed of the track can then be defined in the TAC.

2. TAC

The TAC shows the attributes of a hooked track like ID, environment, IFF, bearing, distance, position, course, speed, altitude.

Operations:

To assign/change an ID:

Click on the scroll-down menu besides the ID → select an ID e.g. hostile → window is closed and ID is changed to hostile

Change the values of NRTT-Tracks:
Type in the coordinates, the course and the speed of the NRTT in the respective text fields (for real-time tracks these values can not be changed)

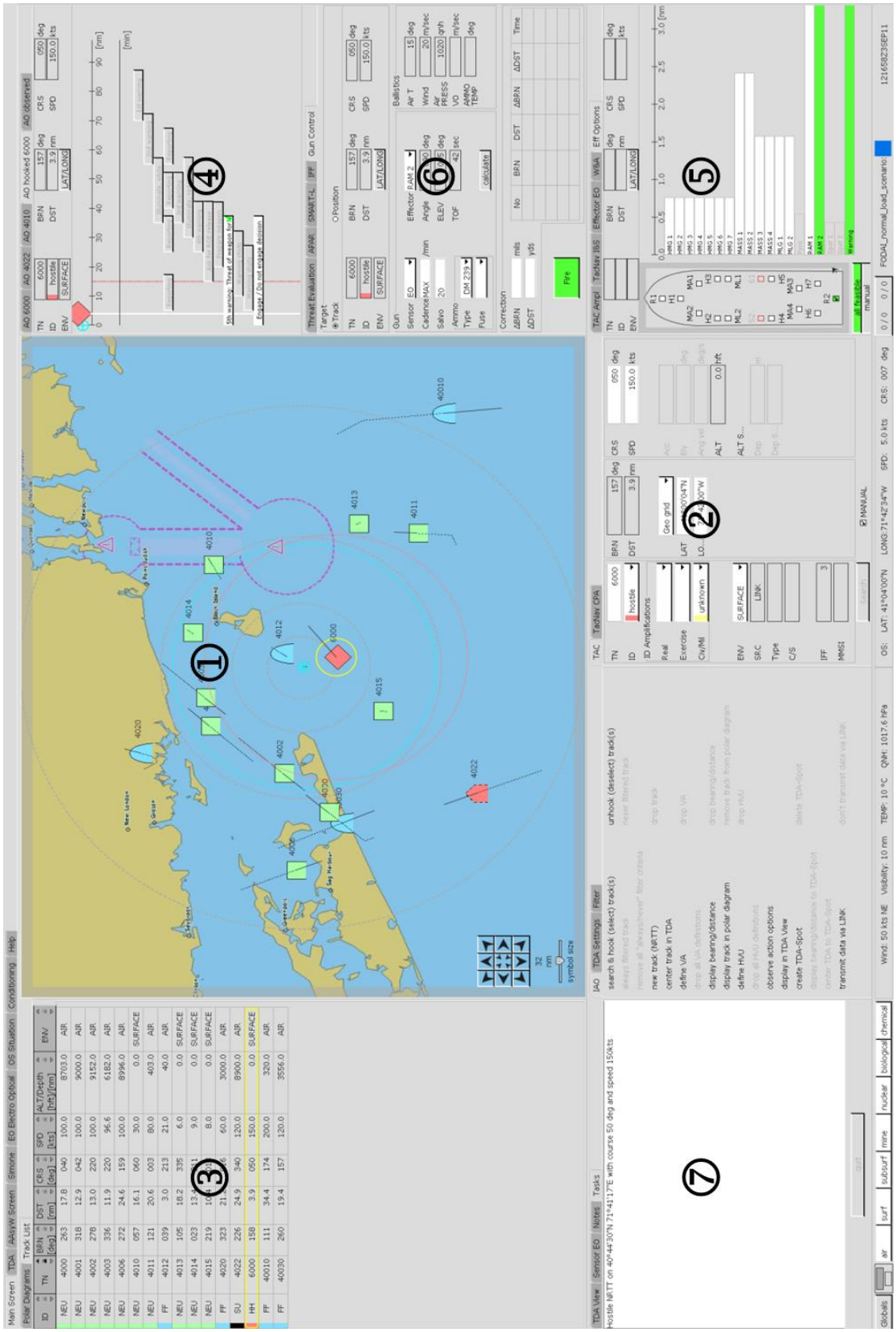


Figure 1: Main screen of the demonstrator

3. Track list

The track list shows all tracks with their most relevant properties like their ID, Bearing, Distance, Course, Speed, Altitude and Environment. The tracks are listed in the order of their track number. However it is also possible to sort them according to other criteria in the header. It can be used for monitoring the properties of the tracks and detecting changes of their behaviour.

4. Warnings and action options

This display area contains several action options which are listed in relation to the distance of a track to the own ship or the time needed by the track to hit the own ship by means of its furthest reaching effector. It is also possible to express warnings when the respective track falls below a certain distance to the own ship

Operations:
Expression of warnings:
Click on the respective warning button e.g. 5th warning: Threat of weapon force → a green check mark appears

5. Effector options

The Effector Options panel is originally located on the screen for engagement in the close-up range and is put on the main screen just for the test. It shows with which effectors the track is currently engagable.
Green = Track is engagable with this effector
Yellow= Effector is currently in use
White = Effector is basically usable but the track is out of reach
Grey = Effector is not usable

The button “All feasible” selects all effectors with which the track is engagable.

Operations:
Selection of an effector for the engagement of a track
Hook the track that is to be engaged → click on one of the effectors that are marked green → The Fire-Button will turn from red to green in the panel Gun Control.

6. Gun Control

The Gun Control panel shows some information on the hooked track which is to be engaged together with information on the selected effector and enables the initiation of an engagement.

Operations:
To engage: Click on the green fire button.

7. Tasks

This panel has been created for the test to inform the operator on NRT-Tracks. If a new notification in this panel appears the operator is requested to create a NRT- track with the coordinates, course and speed given in the note.

C.3 Ergebnisse auf individueller Ebene (Experiment 3)

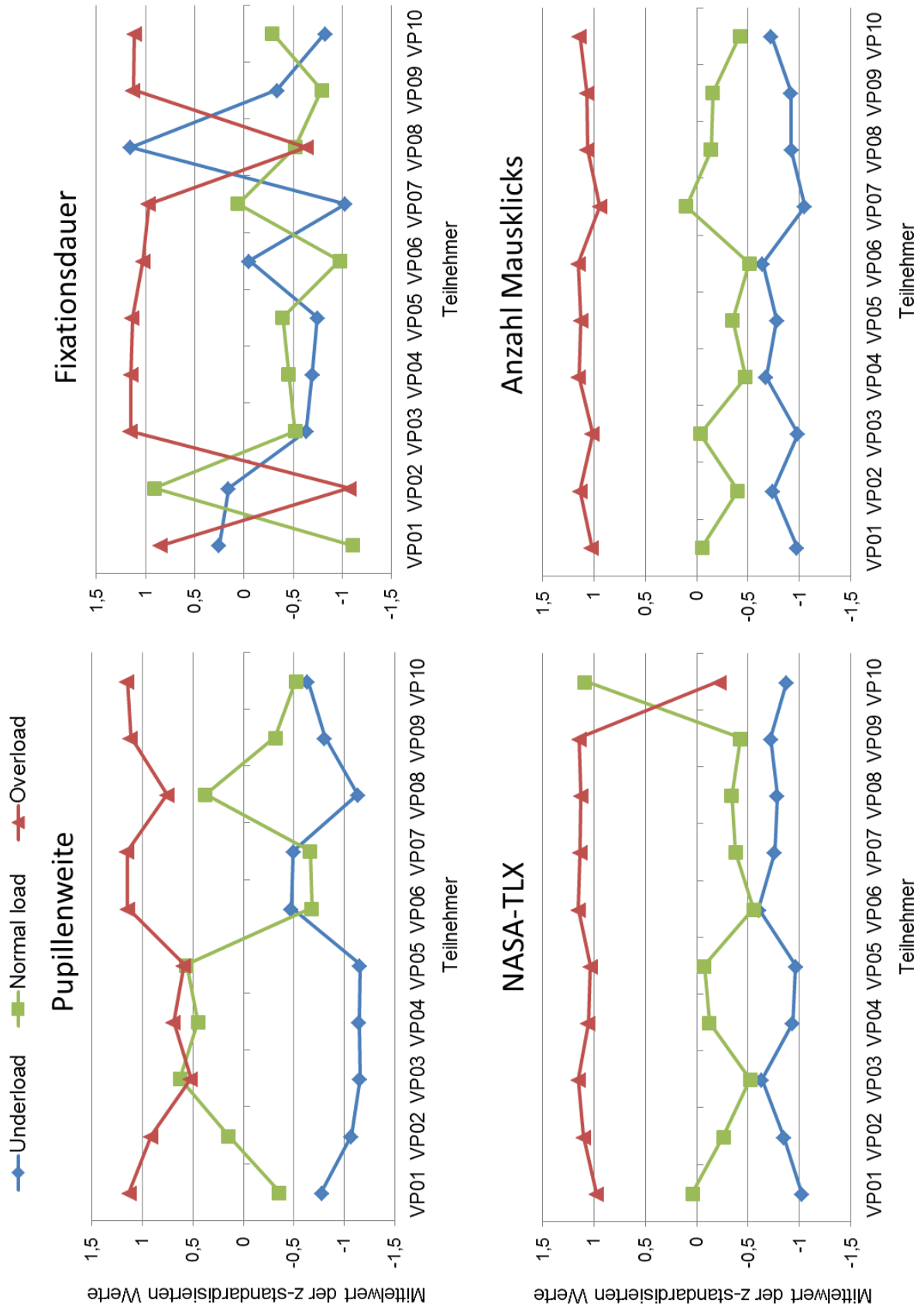
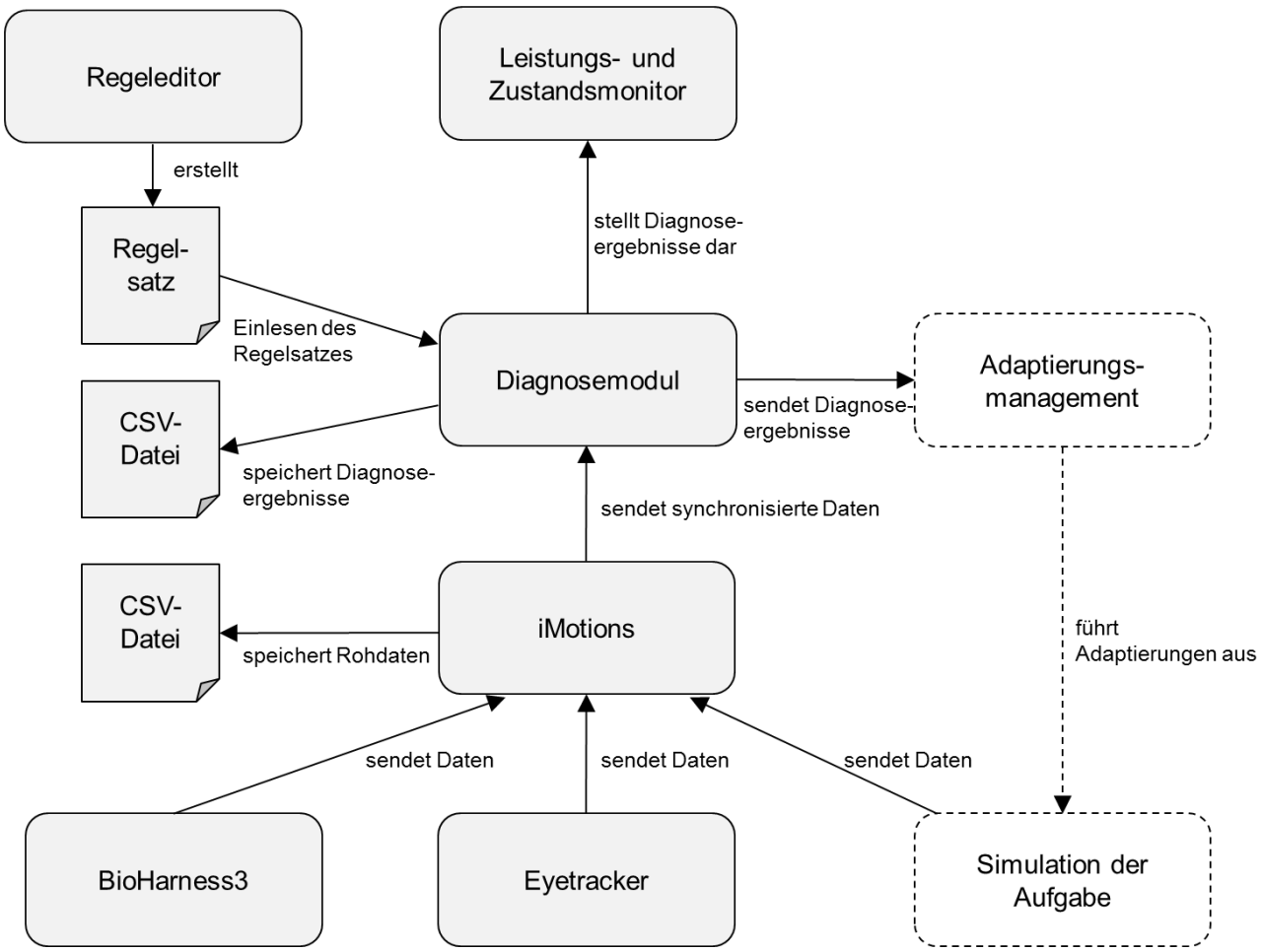


Abbildung 55. Mittelwerte der betrachteten Nutzerzustandsindikatoren in den Versuchsbedingungen Underload, Normal Load und Overload pro Versuchsperson (VP)

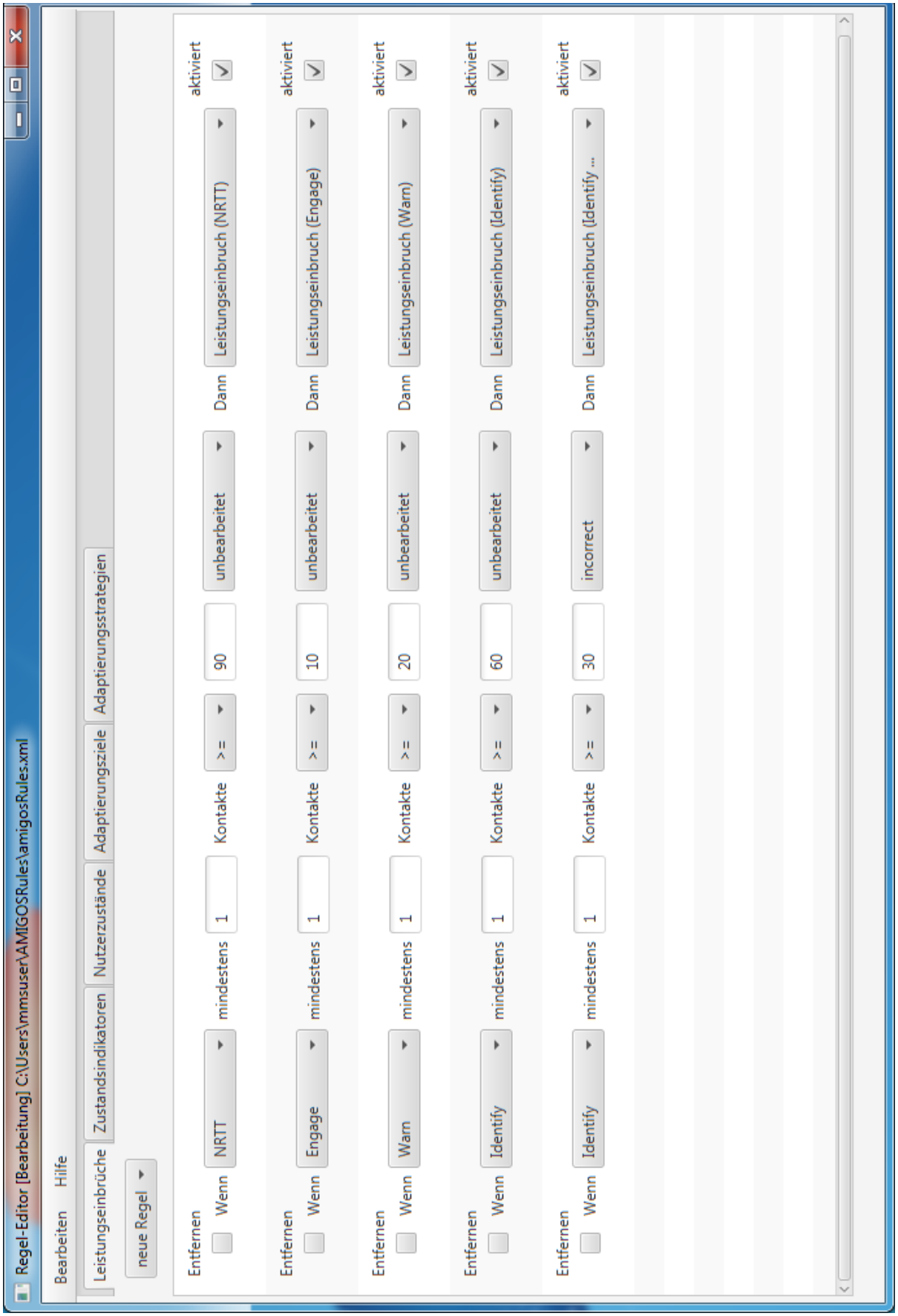
Anhang D. Darstellungen zur technischen Umsetzung der Echtzeitdiagnose

D.1 Architektur der Soft- und Hardwarekomponenten

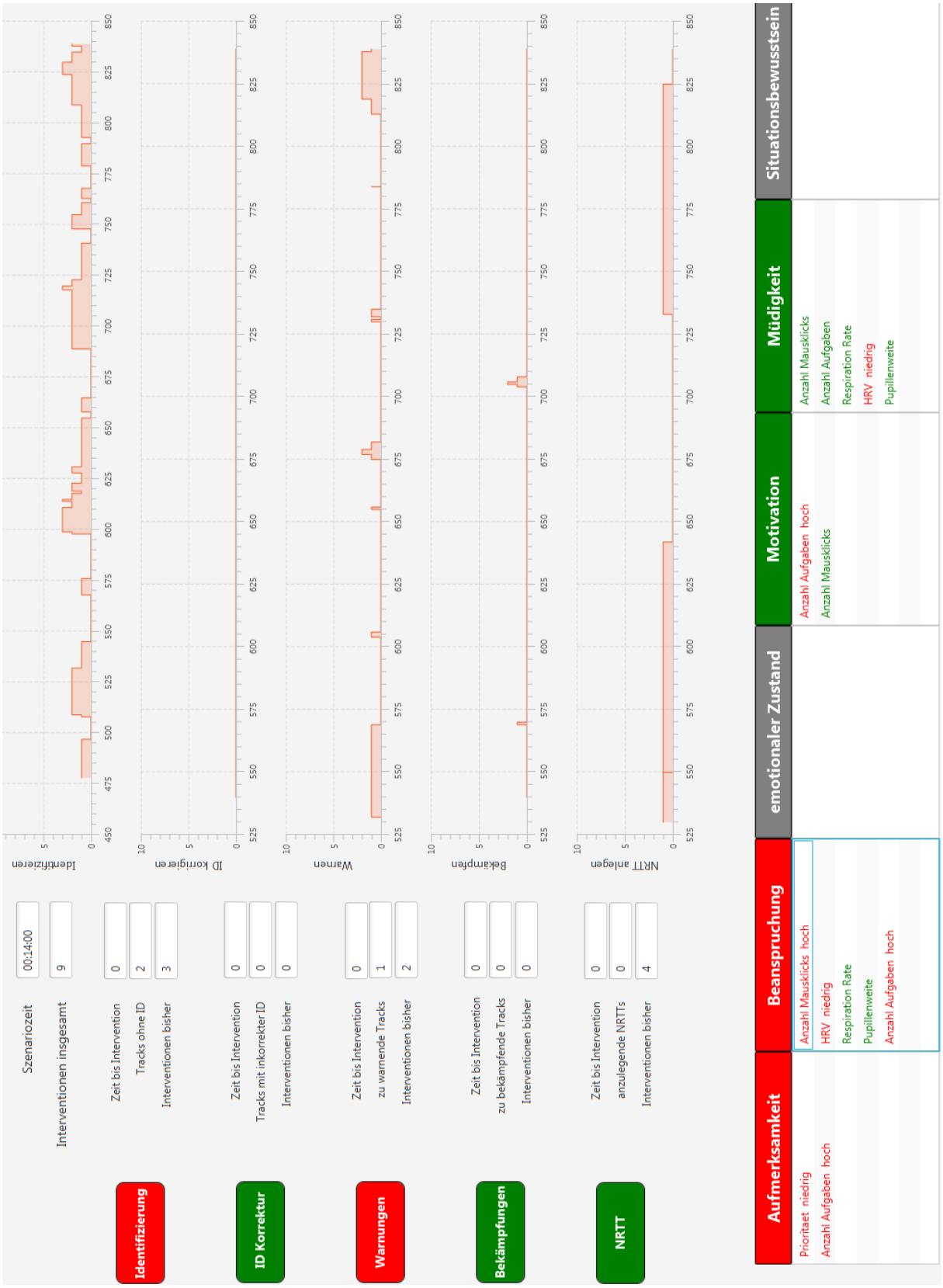
Die Grafik veranschaulicht die Architektur der Soft- und Hardwarekomponenten, die Bestandteil der Echtzeitdiagnose RASMUS sind (vgl. Beschreibung in Abschnitt 6.7). Das Adaptierungsmanagement und die Simulation der Aufgabe sind gestrichelt dargestellt, da sie nicht Bestandteil von RASMUS sind, aber mit RASMUS in direkter Verbindung stehen.



D.2 Regeleditor



D.3 Leistungs- und Zustandsmonitor



Anhang E. Versuchsmaterialien und ergänzende Auswertungen zu Experiment 4

E.1 Instruktion (Experiment 4)

Liebe(r) Versuchsteilnehmer(in),

vielen Dank für Ihre Bereitschaft, an unserem Versuch teilzunehmen!

In diesem Experiment möchten wir untersuchen, ob die von uns entwickelte Methode zur Nutzerzustandsdiagnose in der Lage ist, Zustandsveränderungen beim Menschen, die sich bei der Interaktion mit einem technischen System ergeben, korrekt zu erkennen. Später soll diese Diagnosefunktion dazu genutzt werden, um die Funktionsweise und das Verhalten des technischen Systems (im Sinne eines adaptiven Systems) an den jeweiligen Nutzerzustand anzupassen.

Für die Diagnose von Nutzerzuständen verwenden wir in diesem Experiment als physiologische Sensoren einen Eyetracker und einen Brustgurt. Mit dem Eyetracker werden Blickbewegungen sowie die Pupillenweite erfasst. Mit dem Brustgurt werden verschiedene Vitalfunktionen, wie die Herzrate, die Atmungsrate, die Neigung des Oberkörpers und die Körpertemperatur aufgezeichnet, die als Indikatoren für Nutzerzustände herangezogen werden können.

Der Ablauf der Untersuchung gliedert sich folgendermaßen:

- Einweisung
- Fragebogen zur Person und zum momentanen Befinden
- Anpassen und Kalibrieren der physiologischen Sensoren
- Übungsszenario
- Versuchsdurchführung (ca. 45 Minuten + Unterbrechungen für Fragebögen)

Insgesamt wird die Untersuchungsdauer ca. 2 Stunden betragen.

Viel Spaß und viel Erfolg!

WICHTIG!

Bitte andere Kollegen nicht über die Inhalte des Experiments informieren!
Für uns ist es wichtig, dass die Versuchsteilnehmer den Szenarioverlauf vorher nicht kennen.

Danke!

Beschreibung der Experimentalaufgabe

Bei der Experimentalaufgabe übernehmen Sie die Rolle eines Marineoperateurs, der Luftkontakte („Tracks“) auf einem simulierten Radarbildschirm („Tactical Display Area“, kurz: TDA) überwacht. Die Aufgabe beinhaltet verschiedene Teilaufgaben, die im Folgenden kurz erläutert werden. Wie diese Aufgaben ausgeführt werden können, wird Ihnen während des Übungsszenarios gezeigt.

Für die Aufgaben steht unterschiedlich viel Zeit zur Verfügung bevor diese eine Unterbrechung des Szenarios auslösen.

Identifizierung (60 Sek.)
Noch nicht identifizierte Kontakte erscheinen als gelbe Symbole auf der TDA. Diese müssen anhand von ID-Kriterien (siehe 2. Blatt) als „friend“ (freundlich), „hostile“ (feindlich), oder „neutral“ (zivil) identifiziert werden. Ändern bereits identifizierte Kontakte ihr Verhalten, so dass dies eine Änderung ihrer Identität notwendig macht, müssen diese umidentifiziert werden.

Anlegen von NRT-Tracks (90 Sek.)
Von Zeit zu Zeit müssen sogenannte „Non real-time tracks“ (NRTTs) manuell auf der TDA angelegt werden. Dies wird durch einen Briefumschlag im Fenster links unten auf dem Bildschirm angezeigt. Bei Klick auf den Briefumschlag erscheinen Informationen, die für das Anlegen des NRTT benötigt werden (Identität, Koordinaten, Geschwindigkeit, Höhe, Flugrichtung).

Warnung (20 Sek.)
Kontakte, die „hostile“ also feindlich sind, müssen gewarnt werden, sobald Sie in die Sicherheitszone („Identification Safety Range“, kurz: ISR) des Eigenschiffs eindringen. Diese ist auf der TDA durch einen blauen Kreis kenntlich gemacht.

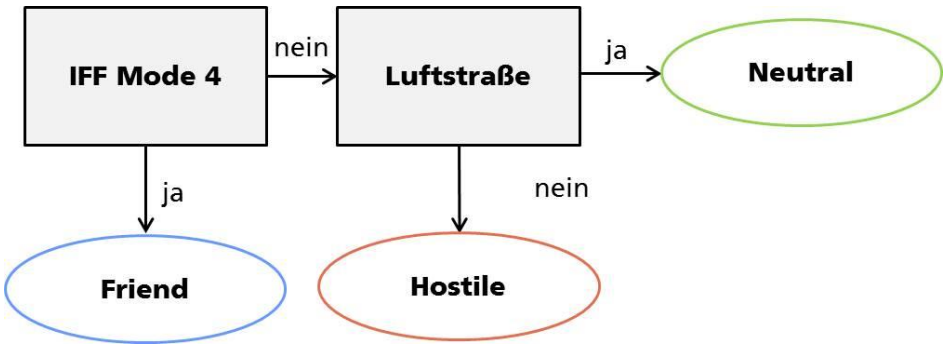
Bekämpfung (10 Sek.)
„hostile“-Kontakte, die trotz Warnung auf das Eigenschiff zufliegen, müssen bekämpft werden, sobald sie in die Weapon Range (WR) eindringen. Die Weapon Range ist auf der TDA durch einen roten Kreis kenntlich gemacht.

ID Kriterien und Priorität der Aufgaben

Kriterien für die Identifizierung:

- IFF-Mode
- Luftstraßenkonformität

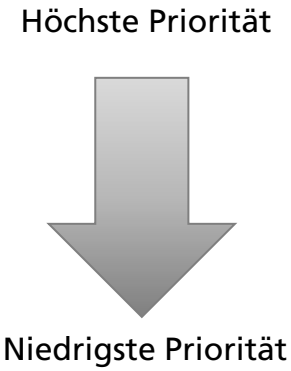
Vorgehen bei der Identifizierung:



Priorität der Aufgaben:

Wenn mehrere Aufgaben zur gleichen Zeit bearbeitet werden müssen, sollte die Aufgaben mit höchster Priorität zuerst bearbeitet werden.

1. Bekämpfung
2. Warnung
3. Identifizieren von Kontakten innerhalb der ISR und WR
4. Anlegen von NRTT
5. Identifizierung von Kontakten außerhalb der ISR und WR



E.2 Versuchsprotokoll (Experiment 4)

Versuchsprotokoll

Datum: _____ Probandennummer: ___

Uhrzeit: _____

Übungsszenario Beginn:_____ Ende:_____

Versuchsdurchgang Beginn:_____ Ende:_____

E.3 Zusatzauswertungen (Experiment 4)

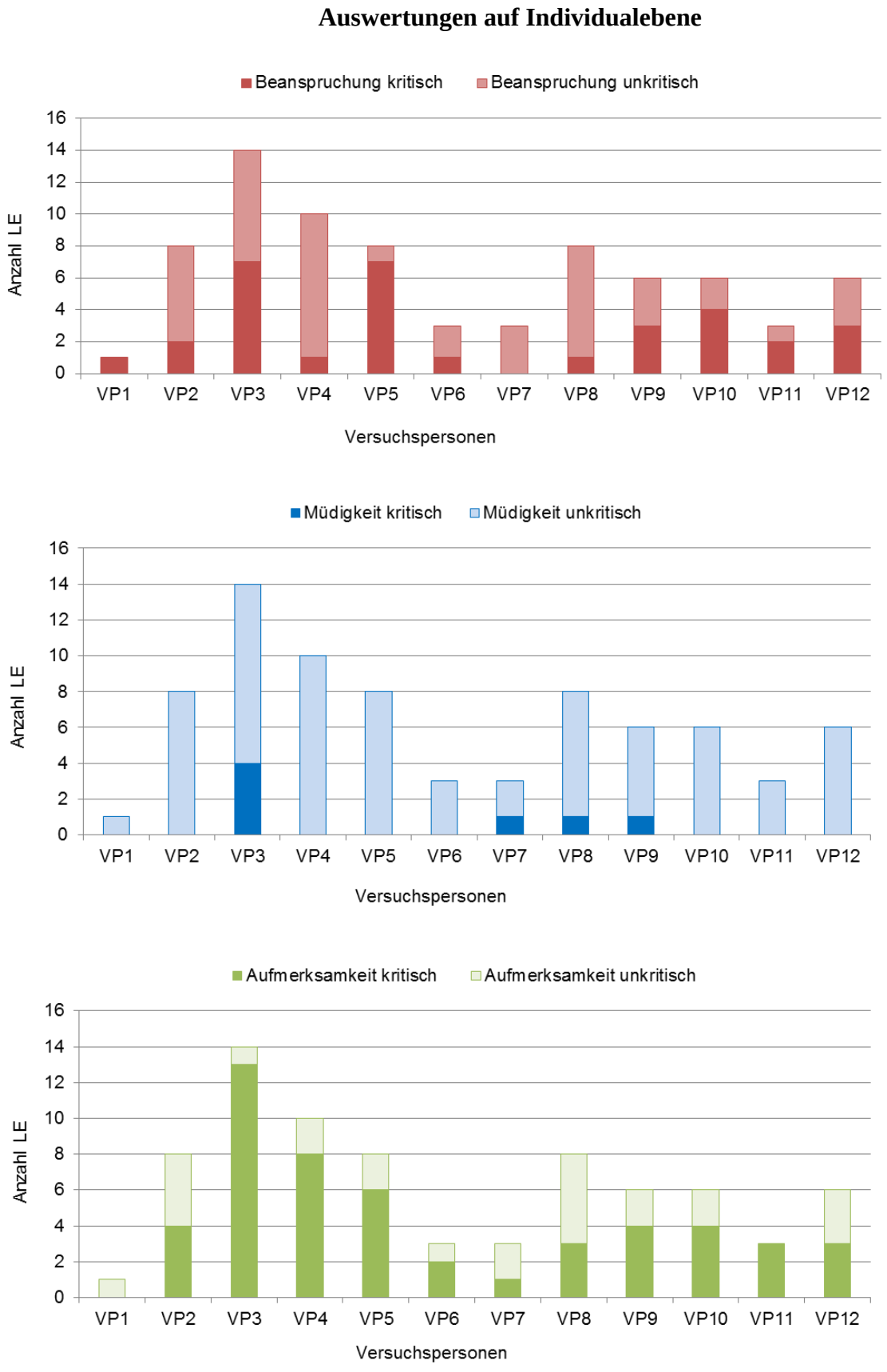
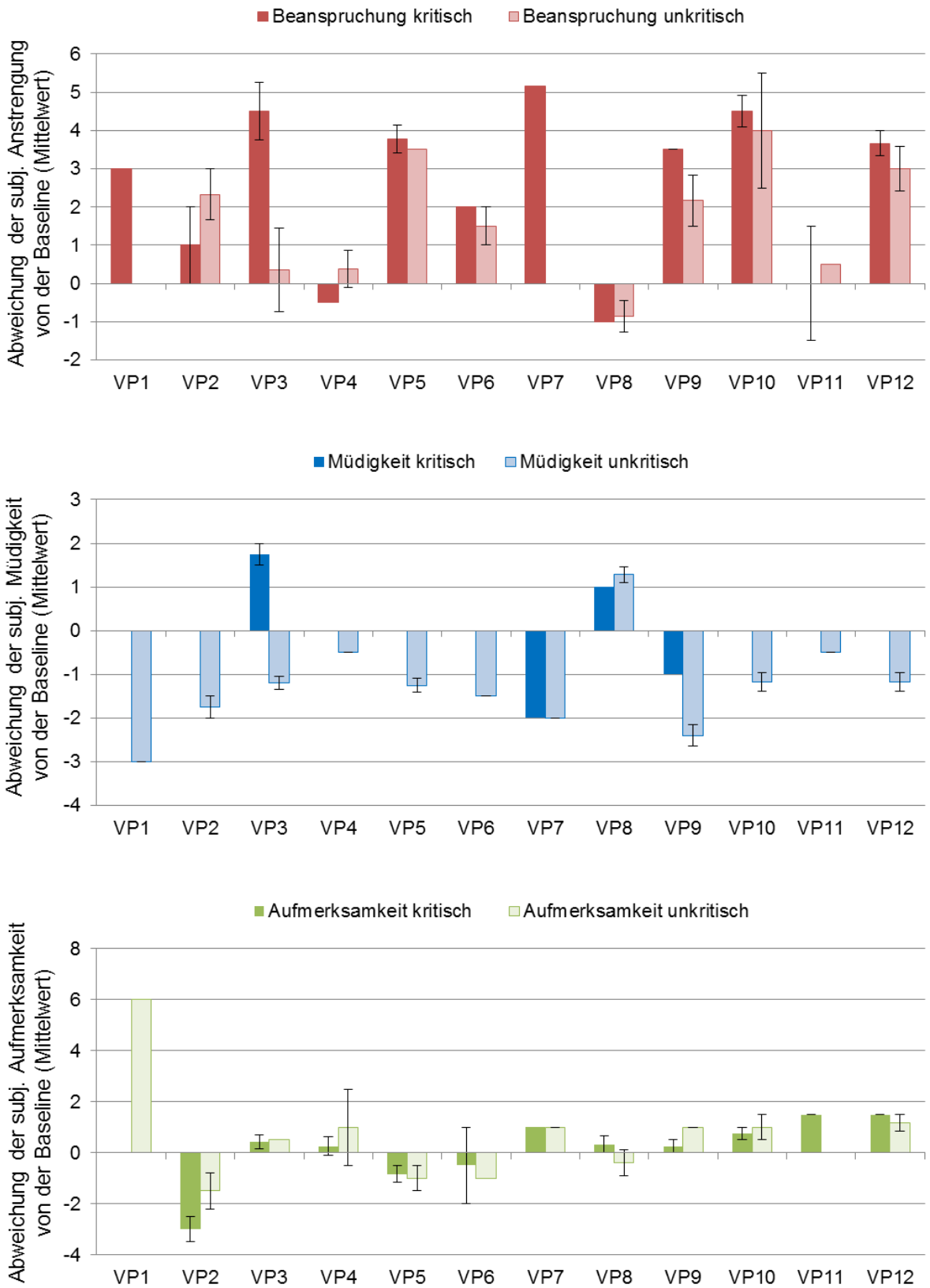
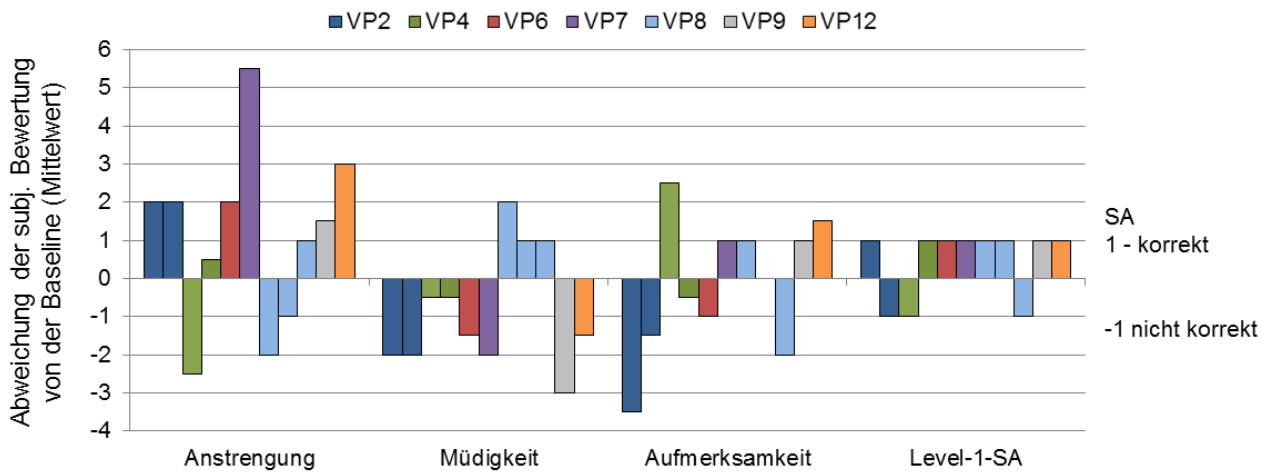


Abbildung 56. Häufigkeit der Leistungseinbrüche (LE) mit (un-)kritischer Beanspruchung, (un-)kritischer Müdigkeit und (un-)kritischer Aufmerksamkeit pro Person



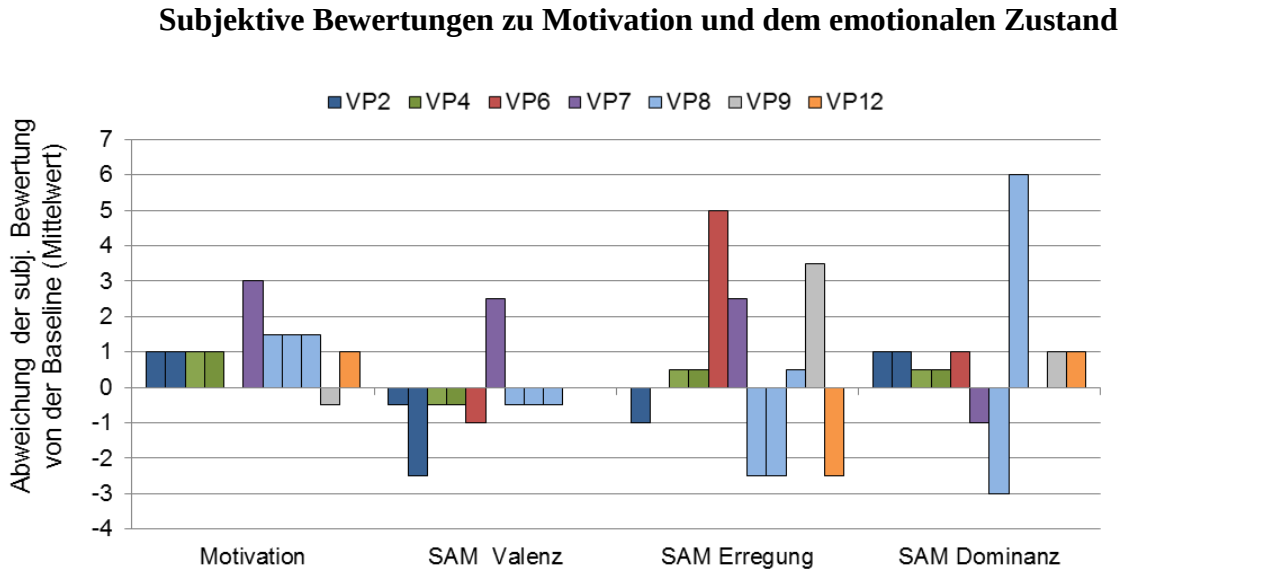
Anmerkung: Zu beachten ist, dass diese Ergebnisse auf unterschiedlichen Fallzahlen pro Person beruhen, die Abbildung 56 entnommen werden können.

Abbildung 57. Durchschnittliche Abweichungen der subjektiven Bewertung von der Baseline bei (un-)kritischer Beanspruchung, (un-)kritischer Müdigkeit und (un-)kritischer Aufmerksamkeit pro Person



Anmerkung: Balken gleicher Farbe beziehen sich auf die gleiche Versuchsperson.

Abbildung 58. Baseline-Abweichungen der subjektiven Bewertungen bezüglich Anstrengung, Müdigkeit und Aufmerksamkeit bei Leistungseinbrüchen ohne kritisch diagnostizierten Nutzerzustand



Anmerkung: Balken gleicher Farbe beziehen sich auf die gleiche Versuchsperson. Bezüglich Valenz weisen hohe Werte auf eine negative Valenz hin, bezüglich Dominanz weisen hohe Werte auf ein Gefühl geringer Dominanz hin.

Abbildung 59. Baseline-Abweichungen der subjektiven Bewertungen bezüglich Motivation und den Dimensionen des SAM bei Leistungseinbrüchen ohne kritisch diagnostizierten Nutzerzustand

Beitrag der Indikatoren zu den Diagnoseergebnissen kritische Beanspruchung und kritische Müdigkeit

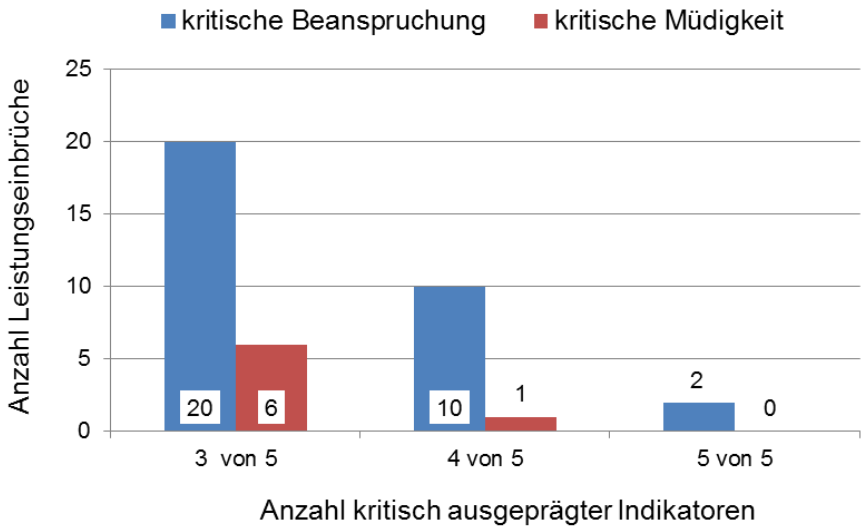


Abbildung 60. Anzahl kritischer Indikatoren bei Leistungseinbrüchen mit kritischer Beanspruchung und kritischer Müdigkeit

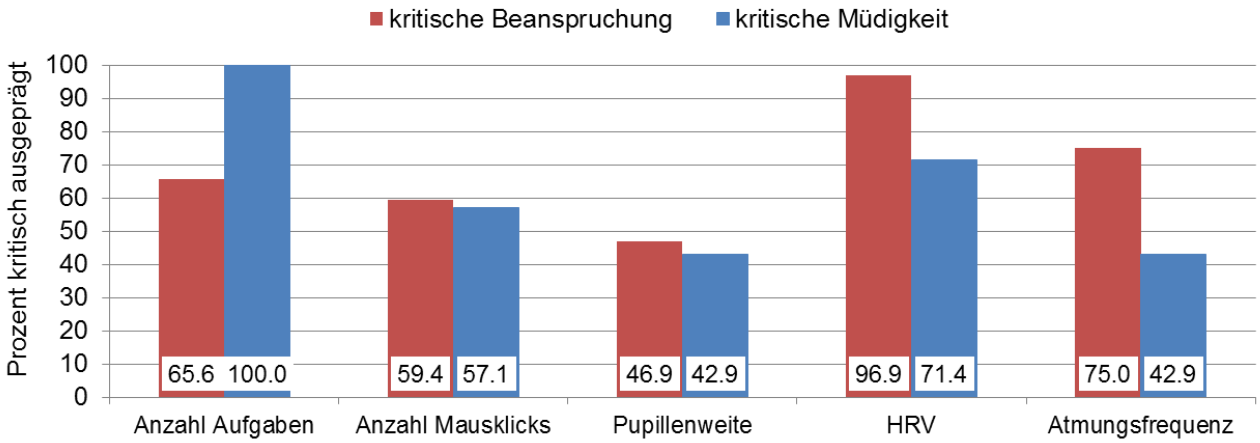


Abbildung 61. Prozentanteile kritischer Indikatorausprägungen bei Leistungseinbrüchen mit kritischer Beanspruchung und kritischer Müdigkeit

Eidesstattliche Erklärung

Hiermit versichere ich schriftlich und eidesstattlich gemäß § 11 Abs. 2 PromO v. 08.02.2011/08.05.2013:

1. Die von mir vorgelegte Dissertation ist selbstständig verfasst und alle in Anspruch genommenen Quellen und Hilfen sind in der Dissertation vermerkt worden.

2. Die von mir eingereichte Dissertation ist weder in der gegenwärtigen noch in einer anderen Fassung an der Technischen Universität Dortmund oder an einer anderen Hochschule im Zusammenhang mit einer stattlichen oder akademischen Prüfung vorgelegt worden.

Wachtberg, den 25.03.2019 _____

3. Weiterhin erkläre ich schriftlich und eidesstattlich, dass mir der „Ratgeber zur Verhinderung von Plagiaten“ und die „Regeln guter wissenschaftlicher Praxis der Technischen Universität Dortmund“ bekannt und von mir in der vorgelegten Dissertation befolgt worden sind.

Wachtberg, den 25.03.2019 _____